DOCUMENT RESUME

ED 094 759 IR 000 959

AUTHOR	Hansen, D. N.; And Others
TITLE	Computer-Based Adaptive Testing Models for the Air Force Technical Training Environment Phase I: Development of a Computerized Measurement System for Air Force Technical Training. AFHRL-TR-74-48.
INSTITUTION	Air Force Human Resources Lab., Lowry AFB, Colo. Technical Training Div.; Florida State Univ., Tallahassee. Computer Applications Lab.
REPORT NO	AFHRL-TR-74-48
PUB DATE	Jul 74
NOTE	81p.; Final report for period 1 January 1973-31 December 1973
EDRS PRICE	MF-$0.75 HC-$4.20 PLUS POSTAGE
DESCRIPTORS	*Computer Programs; Cost Effectiveness; Measurement Techniques; Military Training; *Testing; *Training; Validity
IDENTIFIERS	Air Force; *Computer Managed Instruction; Project PLATO

ABSTRACT

This study explored the utility, from a psychometric and cost effectiveness standpoint, of a computerized adaptive measurement system in an Air Force technical training environment. The phase 1 effort was designed to take the study to the point of producing an operational system ready to actually test technical training students adaptively. A literature review indicated that two testing techniques showed considerable promise: flexilevel testing and hierarchical testing; two courses were selected to implement these procedures. One preliminary conclusion was that adaptive testing appears to offer the potential for time savings of up to 50%. Furthermore, it was found that a very flexible computer system to drive the adaptive testing strategies could be relatively easily developed. (WH)

BEST COPY AVAILABLE

AFHRL-TR-74-48

COMPUTER-BASED ADAPTIVE TESTING MODELS FOR THE AIR FORCE TECHNICAL TRAINING ENVIRONMENT PHASE I:

DEVELOPMENT OF A COMPUTERIZED MEASUREMENT SYSTEM FOR AIR FORCE TECHNICAL TRAINING

By

D. N. Hansen
B. F. Johnson
R. L. Fagan
P. Tam
W. Dick

Florida State University
Tallahassee, Florida 32306

TECHNICAL TRAINING DIVISION
Lowry Air Force Base, Colorado 80230

July 1974
Final Report for Period 1 January 1973 – 31 December 1973

Approved for public release; distribution unlimited.

AIR FORCE HUMAN RESOURCES LABORATORY

ED 094759

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by Florida State University, Tallahassee, Florida 32306, under contract F41609-73-C-0013, Project 1121, with Technical Training Division, Air Force Human Resources Laboratory (AFSC), Lowry Air Force Base, Colorado 80230. Dr. Roger L. Pennell was the contract monitor.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved.

MARTY R. ROCKWAY, Technical Director
Technical Training Division

Approved for publication.

HAROLD E. FISCHER, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFHRL-TR-74-48

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): COMPUTER-BASED ADAPTIVE TESTING MODELS FOR THE AIR FORCE TECHNICAL TRAINING ENVIRONMENT PHASE I: DEVELOPMENT OF A COMPUTERIZED MEASUREMENT SYSTEM FOR AIR FORCE TECHNICAL TRAINING

5. TYPE OF REPORT & PERIOD COVERED: Final, 1 Jan 73 - 31 Dec 73

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): D. N. Hansen, B. F. Johnson, R. L. Fagan, P. Tam, W. Dick

8. CONTRACT OR GRANT NUMBER(s): F41609-73-C-0013

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Florida State University, Tallahassee, Florida 32306

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 1121-03-07

11. CONTROLLING OFFICE NAME AND ADDRESS: HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235

12. REPORT DATE: July 1974

13. NUMBER OF PAGES: 86

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Technical Training Division, Air Force Human Resources Laboratory, Lowry Air Force Base, Colorado 80230

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report): Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

testing	individualized testing	Bayesian testing
computerized testing	tailored testing	programmed testing
adaptive testing	flexilevel testing	automated testing
sequential testing	response contingent testing	multistage testing
branched testing	two-stage testing	affective testing

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Adaptive testing is viewed as a theoretical framework with associated computerized techniques combining to offer solutions to the growing measurement challenges of individualized technical training. It is characterized by three subprocesses: (a) appropriate test selection and entry; (b) tailored presentation of test items; (c) sensitive scoring, diagnosis, interpretation, and reporting. In the context of Air Force technical training, five benefits of adaptive testing are seen: (a) saving instructional and test time; (b) improving reliability and validity of test decisions; (c) optimizing entry and movement within a required

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

BLOCK 19 Cont'd.

criterion-referenced adaptive testing
blocked up-and-down testing
reliability
validity
mastery testing
measurement of technical training

BLOCK 20 Cont'd.

learning hierarchy; (d) assisting training management through refined data specification, collection, and dissemination; and (e) minimizing remediation time. Related literature is reviewed and synthesized into the three areas of subprocesses listed above, including review papers, test selection and student entry, tailored testing with descriptions of eight models, adaptive testing for hierarchical structures, and scoring, diagnosis, interpretation, and reporting. Three proposed studies, based on prior work done in two Air Force technical courses, are described in a design for validation of adaptive testing. These are a flexilevel test study, a hierarchical learning task adaptive test study, and a criterion zone decision study. The paper concludes with discussion and recommendations.

Summary

Problem

The purpose of this study was to explore the utility, from a psychometric and cost effectiveness standpoint, of a computerized adaptive measurement system in an Air Force technical training environment. Considering the uses a computer might be put to in a computer managed instructional system, adaptive testing offers potentially the greatest payoff, since theoretically testing time can be reduced substantially with either an increase in measurement accuracy or no decrease. This, Phase I, effort was designed to take the study to the point of producing an operational system ready to actually test technical training students adaptively. Testing and analyzing the results so obtained will constitute the Phase II effort.

Approach

A thorough review of the literature in the area of adaptive testing was conducted. This review indicated that two testing techniques showed considerable promise: flexilevel testing and heirarchical testing. These procedures were modified by adopting a two-stage approach whereby a student would be branched into the testing net according to a regression estimate of his predicted score. This procedure will hopefully minimize testing time by administering items which are appropriate for the ability level of the examinee. Two courses were selected to implement these procedures; block I of the Precision Measuring Equipment course was selected for heirarchical testing, and block IV of the Inventory Management course was selected for flexilevel testing.

Results

For the two blocks of instruction, a task analysis was performed and appropriate measurement items selected. These items were then incorporated into a computer system for adaptive testing. The testing procedures were programmed in the TUTOR language supported by the PLATO system at the University of Illinois. Three studies were then designed to evaluate the adaptive testing approach: (a) a study to test and validate flexilevel testing, (b) a study to test and evaluate heirarchical testing, (c) a study to explore testing of the examinee in the criterion zone.

Conclusions

The conclusions which can be drawn from the study to date are necessarily preliminary in nature. However, it appears that adaptive testing offers the potential for time savings of up to 50%. Furthermore, it was found that a very flexible computer system to drive the adaptive testing strategies could be relatively easily developed. However, the file handeling and report generation capabilities of the PLATO system, in this phase of development, was found to require considerable ingenuity in programming.

CONTENTS

BEST COPY AVAILABLE

		Page
I.	ADAPTIVE TESTING: AN OVERVIEW	5
1.0	Definition of Adaptive Testing	5
1.1	Role of Adaptive Testing in Air Force Technical Training	6
1.2	Problem Structure	7
II.	REVIEW OF LITERATURE	9
2.0	Background Literature	9
2.1	Literature Reviews	10
2.1.1	Theoretical Reviews	11
2.1.2	Simulation Studies	12
2.1.3	Empirical Studies	12
2.2	Test Selection and Student Entry	21
2.3	Tailored Testing	23
2.3.1	Sequential Item Testing Model	23
2.3.2	Robbins Monro Procedure	24
2.3.3	Branching Models	25
2.3.4	Related Branching Models	26
2.3.5	Hybrid Model	26
2.3.6	Blocked Up and Down Methods	26
2.3.7	Multistage Models	27
2.3.8	Flexilevel Model	27
2.3.9	Summary of Tailored Testing	28
2.4	Adaptive Testing for Hierarchical Learning Structures	30
2.5	Scoring, Diagnosis, Interpretation, and Reports	34
III.	A DESIGN FOR VALIDATION	37
3.0	Implementation and Demonstration of Adaptive Testing: A Design for Validation	37
3.1	Overview	37
3.1.1	Priorities	37
3.1.2	Constraints	37
3.1.3	Demonstration	38
3.2	Air Force Course Analysis and Liaison Activity	38
3.2.1	Item Parameter Analysis	40
3.3	Study One--Flexilevel Validation Study	42
3.3.1	Computer Implementation	42
3.4	Study Two--Hierarchical Learning Assessment	45
3.4.1	Computer Testing Paradigm	45
3.5	Study Three--Criterion Zone Decision Study	47
3.6	Feasibility Results	48
IV.	CONCLUSIONS AND RECOMMENDATIONS	49
4.0	Overview	49
4.1	Entry Processes	49
4.2	Tailoring Testing of Item Presentation	49
4.3	Scoring, Interpretation, and Reporting	49
4.4	Computer Implementation	50
4.5	Recommendations	50

Page

REFERENCES 51
APPENDIX A 59
Computer Program Output. 59
IM Block III Form B Norm-Referenced Analysis Output 61
IM Block III Form B Criterion-Referenced Analysis Output 67
IM Block IV Form A Norm-Referenced Analysis Output 73
IM Block IV Form A Criterion-Referenced Analysis Output 79

I. ADAPTIVE TESTING: AN OVERVIEW

1.0 Definition of Adaptive Testing

In a technical training situation involving large segments of systematic instruction, the frequency with which measurement occurs and the number of measurement items administered constitute a large time demand with respect to the efficiency of training. These time requirements are magnified in individualized, criterion-referenced training situations or in any situation in which the pace at which instruction occurs is determined principally by a trainee's performance on tests (due to the accelerations and remediations effects). This is exactly the situation faced by Air Force trainees for whom the demonstration of mastery on a lesson test is prerequisite to advancement to the next training objective.

Adaptive testing is a theoretical framework with associated computerized techniques that combine to offer solutions to the growing measurement challenges of individualized technical training. Adaptive testing is characterized by three subprocesses: (a) appropriate test selection and student entry, (b) tailored presentation of test items, and (c) sensitive scoring, diagnosis, interpretation, and reporting. For the first process, it is intuitively and empirically obvious that the test or composite test items should be selected to maximize the accuracy and meaningfulness of the outcome decision. In addition, a student should be entered into the test so as to minimize both trivial, easy items and highly difficult or impossibly hard items, while focusing on the presentation of appropriately difficult and discriminating items. Any adaptive test selection and entry process would have to be based on individual student characteristics to be valid.

In turn, the test item presentation should be designed or "tailored" so as to match items to the current performance or ability level of the student. Simply, items that are too easy or too difficult for a student should be avoided. This is the essence of all tailored testing. Real time scoring and individualized movements based on correct/error patterns are major requirements.

Finally, the scoring procedure (right/wrong, average difficulty indices, average of correct item difficulty indices, etc.), the diagnostic interpretation, and the report (quantitative and/or verbal) should be sensitive to all the information on the student. For example, a bright student who is having a "bad day" should be differentially treated from the marginal student who is "all but eliminated." Each stage in this third process of adaptive testing should reflect both individual student data and the requirements of the training system so as to maximize students' learning rates and mastery performance as well as the efficiency of the training system.

In essence, adaptive testing is a more comprehensive measurement model that optimally selects and enters students into the assessment process, tailors the test items, and individually scores, diagnoses, interprets, and reports outcomes to the maximal benefit of the training system. This report now turns to a consideration of how adaptive testing can benefit Air Force technical training.

1.1 Role of Adaptive Testing in Air Force Technical Training

Adaptive testing is a subsystem of an adaptive training system as represented by the Air Force Advanced Instructional System.[1] The conceptual framework allows for a more integrated approach to training and measurement. Five benefits accrue to Air Force technical training from the application of adaptive testing.

First, fairly large chunks of time are devoted to evaluation in order to ascertain whether current training objectives have been mastered, and the trainee can thus advance to the next lesson. In some instances, the ratio of instructional time to testing time may be as high as one hour out of five or six; that is, as much as 16 to 20% of time on task may be devoted to evaluation. This clearly is an inordinately high percentage of time, and it prompts consideration of alternative strategies which would allow reduction of the amount of time spent in assessment, hence maximizing the percentage of time that can be devoted to instruction. Adaptive testing is, first and foremost, cost effective in that it offers a 50% or more reduction in measurement time. As will be revealed in the background literature search, the time savings in improved accuracy and potential acceleration may significantly increase even this time saving by reducing training time.

Secondly, the use of criterion levels for passing tends to magnify measurement errors in the critical decision region. For example, is a student with a score of 89% correct (given a criterion of 90%) really a failure, and does he, therefore, require a retraining cycle? Amplification of this critical decision region would minimize washback and eliminate attrition elements. Adaptive testing improves the precision in the criterion zone in two ways. First, the borderline students can be identified in real time via computer techniques. They can then (a) be given a more discriminating sequential test, or (b) have their wrong answers subjected to a more detailed analysis to determine the degree of partial knowledge. Second, the misleading element of guessing is minimized since the adaptive testing model adjusts item difficulty to the student's performance level; this eliminates the need to guess. Student motivation is maximized by

---

[1] D. N. Hansen, P. F. Merrill, R. D. Tennyson, D. B. Thomas, H. D. Kribs, S. Taylor, and T. G. James, The Analysis and Development of an Adaptive Instructional Model(s) for Individualized Technical Training, Technical Report for Contract No. F33615-71-C-1277, Air Force Systems Command, (Tallahassee; Florida State University, 1973).

matching item difficulty with performance, since this avoids the demoralizing effects of long series of unanswerable items or the tedium of simplistic questions. Thus adaptive testing has the potential to improve the accuracy (reliability) and precision (validity) of the outcome decision.

Third, technical training is replete with numerous learning hierarchies that are structures of interrelated concepts, rules, skills, and subskills. The tension between theory and performance emphases in training reflects these hierarchies in technical training. Moreover, students enter career fields with partial mastery and gaps in their behavioral repertoires. Adaptive testing provides procedures for accurate entry and only appropriate movement within the training-measurement pattern for these hierarchies. The predictive power of the adaptive testing model allows for pretesting and acceleration around mastered subskills. The mixture of practice and testing can be more individualized, and save training time through acceleration or minimized remediation. Thus an adaptive testing model offers an approach to optimal entry and movement within a required learning hierarchy.

Fourth, as technical training becomes more individualized in order to gain improved training time savings, the logistics and information requirements of measurement grow in geometric proportions. An adaptive testing model assists this managerial challenge by specifying essential and only the required student data. The automation by computers improves scoring accuracy, reduces instructor clerical work, and increases availability of information for critical decision making (e.g., elimination). An accrual structure can be built that more accurately predicts future successes and failures. Finally, the adaptive testing model can ultimately be utilized in the diagnostic process so as to minimize remediation time.

Finally, adaptive testing models offer new paradigms for computer utilization within the training process. As general purpose digital computers are being employed for the management and simulation phases of technical training, the addition of the testing function represents a minor increment in computer system cost (i.e., 15% or less increase in cost). To optimally utilize this computing capacity, this research report reflects the goal of synthesizing "state-of-the-art" theoretical testing models into an operational model that fulfills the requirements of individualized Air Force Technical Training.

1.2 Problem Structure

The requirements of this research and development study can be viewed in terms of the subsequent sections of this report. First, an assessment of the "state-of-the-art" in adaptive testing was essential for identifying all feasible approaches and conceptually designing this Air Force adaptive testing model; Section 2 (Background Literature) will describe the results of this search and design process. In turn, the Air Force courses of Inventory Management and Precision Measuring Equipment were analyzed for potential application. Section 3 describes the

results and delineates plans for the validation of the adaptive testing model. This section also describes the computer implementation and demonstration of feasibility. Finally, the conclusion and recommendations will reflect the view of the Florida State University team for future extensions of the adaptive testing model.

II. REVIEW OF LITERATURE

2.0 Background Literature

In assessing the literature which underlies the advancement in individually oriented testing, a complex nomenclature is found to bear on the field. The following list, with citations, gives some concept of this literature:

Adaptive testing[2]
Branched testing[3]
Computer-assisted testing[4]
Computerized testing[5]
Flexilevel testing[6]
Individualized testing[7]
Multistage testing[8]
Programmed testing[9]

---

[2]D. J. Weiss and N. E. Betz, Ability Measurement: Conventional or Adaptive?, Research Report 73-1 Prepared under Contract No. N00014-67-A-0113-0029 NR No. 150-343, Office of Naval Research, (University of Minnesota, 1973).

[3]A. G. Bayroff, Feasibility of a Programmed Testing Machine, Research Study 64-3, (U.S. Army Personnel Office, November, 1964).

[4]J. E. Crick, "A Critical Review of Computer-Assisted Testing" (Unpublished Qualifying Paper, University of Massachusetts, 1972).

[5]D. N. Hansen and G. Schwarz, An Investigation of Computer-Based Science Testing, Institute of Human Learning Technical Report, (Tallahassee: Florida State University, 1968).

[6]F. M. Lord, "The Self-Scoring Flexilevel Test," Journal of Educational Measurement 8, (1971):147-151.

[7]See footnote 2 above.

[8]See footnote 2 above.

[9]T. A. Cleary, R. L. Linn, and D. A. Rock, "An Exploratory Study of Programmed Tests," Educational and Psychological Measurement 28, (1968): 345-360.

Response-contingent testing[10]
Sequential item testing[11]
Two-stage sequential testing[12]

The reason for the profusion of terminology is that there is almost an infinite number of ways of tailoring single test items or adapting blocks of tests to a given individual. While Lord suggested an emphasis on the key feature, namely tailoring the items to the individual, it is the contention of this review that an adaptive approach appears more appropriate.[13] The adaptive testing model and the associated literature review will therefore be primarily organized in three sections, namely: test selection and entry processes; tailored testing; and adaptive scoring, diagnosis, interpretation, and reporting.

Prior to the presentation of these main sections, a concise summary of prior reviews may set the historical framework out of which the project's adaptive testing model grew.

2.1 Literature Reviews

During the past decade there have been numerous reviews of the individualized testing field. Rosenbach's review emphasized the utility approach to sequential testing.[14] The review by Paterson elaborated on the sequential probability decision rules of Wald and their application to ability assessment.[15] [16] Ferguson surveyed in depth the existing

---

[10] R. Wood, "Fully Adaptive Sequential Testing: A Bayesian Procedure for Efficient Ability Measurement," (Unpublished manuscript, University of Chicago, 1972).

[11] D. R. Krathwohl and R. J. Huyser, "The Sequential Item Test (SIT)," American Psychologist 11, (1956):419.

[12] L. J. Cronbach and G. C. Gleser, Psychological Tests and Personnel Decisions (Urbana: University of Illinois Press, 1965).

[13] F. M. Lord, "Some Test Theory for Tailored Testing," In W. H. Holtzman, ed., Computer-Assisted Instruction, Testing and Guidance (New York: Harper and Row, 1970).

[14] J. H. Rosenbach, "An Analysis of the Application of Utility Theory to the Development of Two-Stage Testing Models" (Unpublished Ph.D. dissertation, University of Buffalo, 1961).

[15] J. J. Paterson, "An Evaluation of the Sequential Method of Psychological Testing" (Unpublished Ph.D. dissertation, Michigan State University, 1962).

[16] A. Wald, Sequential Analysis, (New York: Wiley, 1947).

theories and methods of branching testing.[17] Lord reviewed some of the major findings of tailored testing derived from theoretical studies.[18] Bock and Wood, in their survey of test theory for the period from 1966 through 1969, included a section on sequential item testing.[19]

In general, most of these reviews noted the general lack of empirical findings for computer-based testing and made the observation that conventional ability tests tend to provide more accurate measurements than tailoring strategies at the middle or median range of the ability distribution. (It should be noted that all of these latter findings are based on theoretical or simulation studies, and are not consistent with the limited empirical observations.)

Weiss and Betz have provided the most extensive review of adaptive testing to date.[20] They have divided their review into three types of studies: theoretical, simulation, and empirical. This survey concisely summarizes their views in the following paragraphs. A brief summary of their final conclusions, informative as to the focus of their summaries concerning adaptive testing, follows. They consider that adaptive tests are: (1) considerably shorter than conventional tests, with little or no loss in validity or reliability; (2) more reliable than conventional tests in several studies and yielding more nearly constant precision than standard tests throughout the range of abilities; and (3) in several cases more valid, as measured against an external criterion, than are conventional tests.[21]

2.1.1 Theoretical Studies

Weiss and Betz characterize the theoretical studies to date as providing a great deal of comparative information on a variety of test strategies, but yielding limited insight into any inferences to be made for real world context. The rationale for this assertion is based on the fact that all of the empirical studies are concerned only with hypothetical individuals and hypothetical test items. Moreover, these theoretical studies have validities based on a set of highly restricted assumptions (e.g., the probability of a correct response to an item is normally distributed; the discrimination power of all items is constant; items vary

---

[17] R. L. Ferguson, "The Development, Implementation, and Evaluation of a Computer-Assisted Branched Test for a Program of Individually Prescribed Instruction" (Unpublished Ph.D. dissertation, University of Pittsburg, 1969).

[18] See footnote 13 on page 10.

[19] R. D. Bock and R. Wood, "Test Theory," Annual Review of Psychology 22, (1971): 193-223.

[20] See footnote 2 on page 9.

[21] See footnote 2 on page 9 (pages 58-59).

only in difficulty, and all scales are unidimensional in nature). Finally, there are no tests of significance for the information index offered by the theoretical studies and therefore no empirical methods for determining the relative differences among them. For present purposes, the theoretical studies can be viewed as offering a potential road map for selecting the most appropriate models applicable to the technical training area.

2.1.2 Simulation Studies

A number of studies were reviewed that simulated with real or generated data. Table 1 summarizes the more pertinent simulation studies. As in theoretical studies, Weiss and Betz comment: "They can be used simply as a preliminary device for the technical comparisons of certain adaptive strategies, but results should not be considered definitive until they are replicated in empirical live testing studies."[22]

2.1.3 Empirical Studies

The limited number of empirical studies reviewed by Weiss and Betz indicated a number of serious problems, namely, a confusion of testing methods, be this paper and pencil or computer, small samples or careless experimental procedures, etc. (see Table 2). These problems give rise to serious questions regarding the validity of the studies. In spite of the limitations cited, Weiss and Betz make a strong argument for empirical research:

> It is only through empirical studies that the actual effects of adaptive test administration on the testee and his performance will ultimately become known. Future empirical studies of adaptive testing should be based on reasonably large numbers of subjects from carefully defined populations, using tests based on well-structured item pools normed on large and appropriate groups of subjects, with tests pretested to obtain appropriate kinds of score distributions and probably computer-administered to reduce the extraneous sources of variance in test scores.[23]

These methodological remarks will strongly influence the proposed research activities to be described in Section III.

The Hansen, Hedl, and O'Neil review scanned the literature from a different viewpoint, namely, (a) computer-based test administration, (b) scoring, and (c) reports.[24] The review of computer-based test administration indicated a lagging of measurement studies behind the advances in technological capability, particularly in the area of software.

---

[22] See footnote 2 on page 9 (page 44).

[23] See footnote 2 on page 9 (page 43).

[24] D. N. Hansen, J. J. Hedl, and H. F. O'Neil, Review of Automated Testing, Technical Memo No. 20 (Florida State University, 1971).

TABLE 1

SUMMARY OF SELECTED SIMULATION STUDIES

Author(s)	Model(s)	Sample	Evaluation	Results
Paterson[a]	Branched	Simulated subjects	Accuracy of prediction of ability	The branched test gave more precise test scores for the extreme subjects
Cleary,[b] et al.	Two-stage branched, & sequential	4,885 college applicants	Correlations with "parent" test and external criterion	The two-stage procedures appear to have high reliability and validity
Bryson[c]	Branched	100 students	Correlation with "parent" test	Simulation results differed from empirical results
Wood[d]	Bayesian	Simulated subjects	Comparision with empirical results	The Bayesian ability estimates converged at around 20 items

[a]See footnote 15 on page 10.

[b]See footnote 9 on page 9.

[c]R. Bryson, "Shortening Tests: Effects of method used, length and internal consistency on correlation with total score" Proceedings, 80th Annual Convention of the American Psychological Association, (1972): 7-8.

[d]R. Wood, "Computerized Adaptive Sequential Testing" (Unpublished Ph.D. dissertation, University of Chicago, 1971).

TABLE 1 - continued

Author(s)	Model(s)	Sample	Evaluation	Results
Wood[e]	Branched	91 students	Correlation with course grades	Low correlation between the branched test and course grades
Bryson[f]	Branched	513 subjects	Correlation with parent test	Part-whole correlation of branched test same as that of a 5-item conventional test
Ferguson[g]	Wald's sequential sampling	75 primary graders	Correlation with parent test	High test-retest reliability and high correlation with parent test

Note.--This summary serves only as a guide.

[e]R. Wood, "The Efficacy of Tailored Testing," Educational Research 11, (1969): 219-222.

[f]See footnote c on page 13.

[g]See footnote 17 on page 11.

TABLE 2

SUMMARY OF SELECTED EMPIRICAL STUDIES

Author(s)	Model(s)	Sample	Evaluation	Results
Krathwohl & Huyser[a]	Branched	100 college students	Correlation with "parent" test	Part-whole correlation of 0.78 between the sequential test and parent test
Angoff & Huddleston[b]	Two-stage	6,000 college students	Correlation with GPA	Multilevel test more reliable and valid than single conventional test
Bayroff, Thomas, & Anderson[c]	Branched	500 Army trainees	Correlation with parent test	Correlation of 0.63 for the six-item sequential test with the parent test
Bayroff & Seeley[d]	Branched	102 subjects	Correlation with parent test	Correlation in the vicinity of 0.80
Hansen[e]	Branched	56 students	Correlation with conventional test & external criterion	The branched test had higher correlation with the external criterion than the conventional test

Note.--This summary serves only as a guide.

[a]See footnote 11 on page 10.

[b]W. H. Angoff and E. M. Huddleston, The Multilevel Experiment: A Study of a Two-Stage Test System for the College Board Scholastic Aptitude Test, Statistical Report 58-21, (Princeton, N. J.: Educational Testing Service, 1958).

[c]A. G. Bayroff, J. J. Thomas, and A. A. Anderson, Construction of an Experimental Sequential Item Test Research Memorandum 60-1, (Washington, D.C.: U. S. Army Personnel Research Office, January, 1960).

[d]A. G. Bayroff and L. C. Seeley, An Exploratory Study of Branching Tests, Technical Research Note 188, (Washington, D.C.: U.S. Army Behavioral Science Laboratory, 1967).

[e]D. N. Hansen, "An Investigation of Computer-Based Science Testing," in R. C. Atkinson and H. A. Wilson, (Eds.), Computer-Assisted Instruction: A Book of Readings (New York: Academic Press, 1969).

Computer-based test administration may be described in terms of four areas of methodological activity: (a) terminal equipment, (b) the interactive testing process, (c) reliability and validity issues, and (d) the collection of multiple response indices.

It is now possible to find typewriters, cathode ray tubes, and slide projectors being used for test item presentation. Since the creation of inexpensive terminal equipment is one of the dynamic areas in computer technology, one can anticipate more sophisticated terminal devices as well as significant decrease in the cost. On the other hand, progress with respect to the operation of appropriate audio presentation units and natural speech analyzers has been discouraging. Although digitalized speech as well as speech analysis devices are being investigated at Stanford and Haskins Laboratories respectively, the generic problems involved in natural speech analysis are delaying developments of new equipment. In regard to psychomotor/manipulative presentations, cost seems to be one of the greatest deterrents to any extensive development. Therefore, studies noted will focus on the cognitive/symbolic aspects of adaptive testing.

Turning to the characteristics of the student-terminal interaction, several investigators have provided indirect evidence that this man-machine dialogue may be characterized as unbiased, nonstressful, and personalized in nature. For example, Smith points to a "confession machine effect" which appears to enhance the data acquisition in particular content areas such as the subject's personal experience or his perceived personality characteristics.[25] Evans and Miller found that students responded with greater honesty and candor to highly personal items of a social science questionnaire, and Cogswell and Estavan have reported similar findings on the apparent confidentiality of the computer interview.[26] [27] Therefore, the feasibility of using adaptive testing techniques on the student course critique appears promising.

Evidence for the nonthreatening nature of a computer-based evaluation comes from a study by Gallagher.[28] He investigated the relationship

---

[25] R. E. Smith, "Examination by Computer," Behavioral Science 8, (1963): 76-79.

[26] J. M. Evans and J. R. Miller, "Differential Effects on Response Bias of Computer vs. Conventional Administration of a Social Science Questionnaire: An Exploratory Methodological Experiment," Behavioral Science 14(3), (1969): 216-227.

[27] J. F. Cogswell and D. P. Estavan, Explorations in Computer-Assisted Counseling, TM-2582, (System Development Corporation, 1965).

[28] P. D. Gallagher, An Investigation of Instructional Treatments and Learner Characteristics in a Computer-Managed Instruction Course, Technical Report No. 12, (Tallahassee: Florida State University, CAI Center, 1970).

of instructional treatments and learner characteristics in a terminal-oriented computer-managed instruction course. Computer evaluation and instructor evaluation of term projects produced performance scores which were negatively related to trait anxiety ($r = -.51$) in the instructor-evaluated group, but were not related in the computer-evaluated group ($r = -.03$). One might assume that the treatment group which emphasized human interaction resulted in a greater threat to the individual's self-esteem.

Cronbach cites a number of advantages of computerized tailored testing, namely, excellence of standardization, control of bias, precision of timing, and the integration of learning and testing.[29]

Reliability and validity studies concerning automated administration procedures have demonstrated, from an empirical standpoint, the feasibility of a technological approach, and have paved the way for further research and development efforts. For example, Elwood developed a noncomputerized automated testing booth to administer the Wechsler Adult Intelligence Scale (WAIS).[30] Orr reported favorable results for this approach from a comparison of an automated WAIS presentation with a traditional WAIS presentation ($r = .83$). However, this system only provides scoring capabilities for 2 of the 11 subtests (Digit Span and Digit Symbol).[31] Recent computer methodology describes how the administration of intelligence test items can be programmed to allow for repetition and expansion of verbal responses.[32] This more contingent, interactive elicitation of responses yields equivalent (slightly superior) reliability and validity indices to those found for human presentation. This demonstrated the objective facets of computer-based testing.

In a study of computer-based branched testing, Hansen found a significant improvement in internal consistency reliability for computer presentation ($r = .80$) in comparison with a conventional classroom

---

[29] L. J. Cronbach, Essentials of Psychological Testing (3rd ed.; New York: Harper and Row, 1970).

[30] D. L. Elwood, "Automation of Psychological Testing," American Psychologist 24(3), (1969): 287-289.

[31] T. B. Orr, "A Comparision of the Automated Method and the Face-to-Face Method of Administering the Wechsler Adult Intelligence Scale," paper presented at the meeting of the Indiana Psychological Association, Indianapolis, April, 1969.

[32] J. J. Hedl, Jr., An Evaluation of a Computer-Based Intelligence Test, Technical Report 21, (Tallahassee: Florida State University, CAI Center, 1971).

achievement test (r = .43).[33] More interestingly, the computer-based test yielded a significant relationship (r = .76) with a college entrance aptitude score. In addition, Hansen found that the addition of subjective confidence responses yielded improved validity coefficients. Massengill and Shuford have reported similar results.[34]

Obviously, the full potential of multiple dependent measures remains to be empirically explored within automated testing. Multiple dependent measures such as latency, subjective confidence, and anxiety can be incorporated to improve both the diagnostic power and efficiency of the psychometric instruments. Research with the Minnesota Multiphasic Personality Inventory (MMPI) has shown that the information processing time (latency) for a given item is partially a function of the number of characters in the item, the ambiguity of the item, and the social desirability value of the item.[35] Massengill and Shuford have shown that subjective confidence ratings significantly increase test reliability.[36] Hansen reported an improved predictive relationship for a college entrance aptitude measure if confidence scores are included with the right/wrong CAI scores.[37]

Although the employment of computers to calculate test scores and to carry out statistical analyses and summaries of test data has been common for many years, the volume has been growing at a considerable rate. Woods presents a comprehensive survey of the general uses of such data processing techniques in school testing programs.[38] However, the application of these response analysis techniques to online terminal-oriented computer testing systems is a recent advance. We turn now to the consideration of the use of natural language processing for test responses.

---

[33] D. N. Hansen, "An Investigation of Computer-Based Science Testing," in R. C. Atkinson and H. A. Wilson, eds., Computer-Assisted Instruction: A Book of Readings (New York: Academic Press, 1969).

[34] H. E. Massengill and E. A. Schuford, Report on the Effect of "Degree of Confidence" in Student Testing (Lexington, Mass.: The Schuford-Massengill Corporation, 1967).

[35] T. G. Dunn, R. E. Lushene, and H. F. O'Neil, "A Complete Automation of the Minnesota Multiphasic Personality Inventory and a Study of its Response Latencies," paper presented at the annual meeting of the American Educational Research Association, New York City, 1971.

[36] See footnote 34 above.

[37] See footnote 33 above.

[38] E. M. Woods, "Recent Applications of Computer Technology to School Testing Programs," Review of Educational Research 40(4), (1970): 525-539.

Research focusing on the computer aspects centering around input and output of natural language during online communication between the student and the system has been reported by Starkweather; Colby, Watt, and Gilbert; and Weizenbaum.[39] [40] [41] These authors have developed computer techniques to conduct psychotherapeutic dialogues with patients. Hedl, O'Neil, and Hansen have shown that an interactive dialogue is possible with the automated administration of an individualized intelligence test.[42]

Peck and Veldman of the University of Texas have been developing a computer-based system for presenting and scoring responses to a sentence completion test.[43] The problems of syntax were reduced due to the restriction on the subject to use a single word in responding to each sentence stem. The most recent system produces 40 scores from a 36-item form and employs a complex word-root data reduction system.[44] This prototypic tailored inquiry method offers many of the benefits of a traditional interview, and might serve as a basis of future programs which could conduct intensive assessment interviews.

---

[39] J. A. Starkweather, "COMPUTEST, a Computer Language of Individualized Testing, Instruction, and Interviewing," Psychological Reports 17, (1965): 227-237.

[40] M. C. Colby, J. B. Watt, and J. P. Gilbert, "A Computer Method of Psychotherapy: Preliminary Communication," Journal of Nervous and Mental Disease 142(2), (1966): 148-152.

[41] J. Weizenbaum, "ELIZA-A Computer Program for the Study of Natural Language Communication Between Man and Machine." Communications of the Association for Computing Machinery 9, (1966): 36-45.

[42] J. J. Hedl, H. F. O'Neil, and D. N. Hansen, "Computer-Based Intelligence Testing," paper presented at the annual meeting of the American Educational Research Association, New York City, February, 1971.

[43] R. F. Peck and D. J. Veldman, An Approach to Psychological Assessment by Computer, Research Memorandum No. 10, (Austin: University of Texas, 1961).

[44] D. J. Veldman, "Computer-Based Sentence Completion Interviews," Journal of Counseling Psychology 14(2), (1967): 153-157.

Recently, Archambault developed a computerized program to score verbal responses to three of the seven subtests of the Torrance Tests of Creative Thinking.[45] Subject responses to each of the subtests are scored for fluency, flexibility, and originality. Archambault's data indicated that creativity, as defined by Torrance, was judged accurately by a computer. The syntax problems were reduced by analyzing only the frequency of word usage. However, this frequency word usage or word phrase lookup procedure produced significant correlations ranging from .52 to .99 between the computer and the pooled scores of four trained judges. It appears that the use of a computer to score open-ended responses to standardized test items is feasible and should be further investigated.

In reviewing the recent research on the automated interpretation of test results, Hansen, Hedl, and O'Neil pointed out that the challenge facing such automation is the conversion of quantitative indices or profiles into meaningful verbal statements. The main thrust of research in this area has been in the personality rather than the aptitude domain. Thus a number of studies have concentrated on computerized interpretation of MMPI and Rorschach profiles.[46] [47] [48] [49] [50] [51]

---

[45] F. X. Archambault, "A Computerized Approach to Scoring Verbal Responses to the Torrance Tests of Creative Thinking," paper presented at the meeting of the American Educational Research Association, Minneapolis, Marc., 1970.

[46] H. P. Rome, W. M. Swenson, P. Mataya, E. E. McCarthy, J. S. Pearson, and R. F. Keating, "Symposium on Automation Technics in Personality Assessment," Proceedings of the Mayo Clinic 37, (1962): 61-82.

[47] B. C. Gleuck and M. Reznikoff, "Comparison of Computer-Derived Personality Profile and Projective Psychological Test Findings," American Journal of Psychiatry 121(7), (1965): 1156-1161.

[48] J. C. Finney, "Methodological Problems in Programmed Composition of Psychological Test Reports," Behavioral Science 12, (1967): 142-152.

[49] R. D. Fowler, "The Current Status of a Computer Interpretation of Psychological Tests," American Journal of Psychiatry 125(7) Supp., (1969): 21-27.

[50] B. Kleinmuntz, "Personality Test Interpretation by Digital Computer," Science 139, (1963): 416-418.

[51] Z. A. Piotrowski, "Digital Computer Interpretation of Inkblot Test Data," The Psychiatric Quarterly 38, (1964): 1-26.

Relevant to the present concern, however, are the few studies that have dealt with computerized interpretation of aptitude or achievement tests. In one study, Helm programmed the evaluation of a battery of individual scores for each student.[52] The output was designed mainly to direct translation scores, although there was limited capability for comparision and contrast of profile scores. A more innovative development was a program developed by Cogswell and Estavan.[53] This program was designed to evaluate student folders containing information such as grades, aptitude test scores, etc. Agreement between computer statements and the evaluative statements of two counselors was 75%.

The diagnostic nature of the statements from the Cogswell and Estavan program is an important advance in research on automated score interpretation. An automated diagnostic system with interpretive capabilities could be designed to relate instructor strategies to particular student profiles. The system could be designed to look at both academic and personality variables suggesting strategies on a realtime basis.

Another important aspect of an automated diagnostic and interpretive system is the capability for differential interpretive reporting according to the intended audience. Such a system is able to provide, at one time, diagnostic information statements meaningful to a course instructor, and at another time, more sophisticated information for professionals engaged in research activities.

2.2 Test Selection and Student Entry

As can be inferred from the prior reviews, the area of computer selected and/or composed tests is practically nonexistent. Wood reviewed the techniques for computer-composed tests.[54] The Naval CMI project at Memphis illustrates how students can be routed to specific tests.[55] Adaptive selection of tests remains a highly promising topic for future research. Rasch provides a model that yields equivalent individual measurement (scores) from sets of items varying in difficulty.[56] Masang

---

[52] C. E. Helm, "Simulation Models for Psychometric Theories," In Proceedings of American Federation of Information Processing Societies, Vol. 27, Part 1, (Washington, D.C.: Spartan Books, 1965).

[53] See footnote 27 on page 16.

[54] R. Wood, "Computerized Adaptive Sequential Testing" (Unpublished doctoral dissertation, University of Chicago, 1971).

[55] L. G. Harding, C. A. Johnson, and P. A. Salop, An Evaluation of the Use of Chemically Treated Answer Sheets, (San Diego, Calif.: Naval Personnel and Training Research Laboratory, 1973).

[56] G. Rasch, Probabilistic Models for Some Intelligence and Attainment Tests (Copenhagen: Denmark Paedogogische Institut, 1969).

proposed a procedure for item weighting to achieve invariance of test scores under varying test difficulty levels.[57] Obviously, a large storage capacity, general purpose computer allows for the composition of tests in real time, a near infinite solution to the problem.

In turn, adaptive entry of a student into a test arranged in a difficulty hierarchy remains unexplored. Owen has developed a procedure for applying Bayesian concepts to either the appropriate determination of a test or for the tailoring of test items to each student, the methodology being appropriate for each problem.[58] The Bayesian models offer a number of distinct advantages:

1. The step size of difficulty between tests can be of the examiner's choice.

2. The choice of entry is dependent upon previously collected data on each student.

3. The choice of scoring method is less important and is primarily governed by the choice of a loss function selected by the examiner.

4. All of the test item parameters are permitted to vary.

The unfortunate restrictive assumptions concern the unidimensionality of the ability (performance) and independence of responses found in all other tailor-like models. In determining the test, it should be chosen such that the test item parameters lead to minimized a posteriori variance of the ability. As the test proceeds, the posteriori ability of the parameter can be calculated and new estimates of the student's mean ability and variance can be computed with appropriate adjustments within the test as it proceeds. It should be pointed out that there has been little theoretical or theoretical-comparative work (e.g., theoretical simulated comparisons), and no empirical work using this approach. In essence, it appears to be on the very forefront of the state-of-the-art. As very large computing systems become available, Bayesian models should be investigated in terms of their potential primarily for determining test selection, and in addition as perhaps being the most appropriate way of tailoring item presentations to a student.

In turn, adaptive entry of a student into a test arranged in a difficulty hierarchy remains unexplored. In a more integrated instructional and testing paradigm, Suppes has provided for individualized entry for well over 50,000 students in a mathematics CAI drill and practice

---

[57] B. Masang, "Item Weighting: An Approach to Invariance of Test Scores Under Varying Test Difficulty Levels" (Unpublished Preliminary Paper, Florida State University, 1972).

[58] R. J. Owen, A Bayesian Approach to Tailored Testing, Research Bulletin 69-72, (Princeton, N.J.: Educational Testing Service, 1969).

program.[59] The results indicate that students can be given appropriate entry based on the single variable of grade level and find an appropriate performance level within a minimum of one hour of instruction. Results similar to this have been reported by working with a similar public school population.[60] It should be observed that each of these programs utilized only one variable (grade level) for the predicted entry placement. If multivariate regression techniques were utilized, it would undoubtedly be true that a much more precise placement could be determined. It should be observed, though, that the evaluation of placement for adaptive testing will have to be determined in terms of the criterion of minimum number of test item presentations, since the behavioral evaluation is elusive at best, and perhaps impossible to answer in terms of student self-ratings.

2.3 Tailored Testing

In this section, eight different formal models will be presented which provide for a form of matching the item presentation to the ability (performance) indices of the student. For each of the models, a brief characterization and an elaboration of their advantages and limitations will be presented. The formal characteristics of each of the models can be found by searching the literature, and studying its axiomatic and psychometric characteristics.

2.3.1 Sequential Item Testing Model

Arising from statistical decision theory and the sequential probability ratio test developed by Wald, the sequential item testing model establishes three decision outcome spaces: (a) success, (b) failure, and (c) a further test area.[61] As test response data are collected from a student, the sequential analysis is performed and appropriate statistical inference established such that testing continues if the summed statistics remain in the indeterminate stage, and stops if the student is classified in either the success or failure outcome spaces. The earliest descriptions

---

[59] P. Suppes, M. Jerman, and D. Brian, Computer-Assisted Instruction: Stanford's 1965-66 Arithmetic Program. New York: Academic Press, 1968.

[60] D. N. Hansen, B. F. Johnson, E. P. Durall, B. Lavin, and L. McCune, A Rural County Computer-Related Instructional Technology Project, USDHEW Title III Final Report, (Tallahassee: Florida State University, 1971).

[61] See footnote 16 on page 10.

of this model can be found in the early 1950's and its more recent application has been reviewed and reported by Ferguson.[62] [63]

In general, this procedure requires approximately one half the number of test items as the conventional procedure. Its primary advantage is in its efficiency in reaching a decision. Perhaps its most desirable feature is for borderline students, who are given every possible opportunity to pass until all test items are exhausted. The administration of additional items to the borderline student increases measurement accuracy and should be considered a desirable feature to be investigated in an empirical fashion. Unfortunately, however, the model assumes four parameters; that is, the pass and failure boundary points, such as pass at .90 and fail at .85, and the risk factors referred to as alpha and beta error types. Since there is no given rationale for specifying these values, only broad empirical study will provide a basis from which an appropriate selection of parameter values can be derived. Thus, in nature, this will undoubtedly be one of the models which will be implemented in some ultimate and concluding phase of adaptive testing research.

2.3.2 Robbins Monro Procedure

This model, like all of the tailored testing models to be reviewed, starts each student in a median difficulty level and successively reduces the stepsize between item difficulty as the test proceeds. For example, the step size might start out at .20 and successively come down to .01. Testing is continued until the student reaches some difficulty level at which he answers half of the items correctly and half of the items incorrectly. The procedure was created by Robbins and Monro and reviewed by Wetherill and Lord.[64] [65] [66] [67] Stocking also provides an evaluation of the process

---

[62] A. Anastasi, "An Empirical Study of the Applicability of Sequential Analysis of Item Selection," Educational and Psychological Measurement 13, (1953): 3-13.

[63] See footnote 17 on page 11.

[64] H. Robbins and S. Monro, "A Stochastic Approximation Method," The Annals of Mathematical Statistics 22, (1951): 400-407.

[65] G. B. Wetherill and H. Levitt, "Sequential Estimation of Points on a Psychometric Function," British Journal of Mathematical and Statistical Psychology 18, (1965): 1-10.

[66] See footnote 13 on page 10.

[67] F. M. Lord, "Robbins Monro Procedures for Tailored Testing," Educational and Psychological Measurement 31, (1971): 3-31.

as a testing technique.[68]

The Robbins Monro process provides excellent measurement at high and low performance levels, but unfortunately is less efficient than conventional testing in the median range. In addition, the large number of test items required also makes it a burdensome model to implement. The difficulty of implementing the model can be directly related to the observation that no empirical study has been attempted for this process at this date.

2.3.3 Branching Models

This model routes a student through a large network of test items according to a simplistic rule: if a student answers an item correctly, present a slightly more difficult next item; or, if he responds incorrectly, make the next succeeding item easier. The fixed step size is usually set at somewhere between .025 and .05. The model has been described and utilized by a number of investigators.[69] [70] [71]

The primary advantage of the branching model is its improved measurement accuracy at the extremes of the performance continuum, although the complement is also true; namely, poorer performance in the median range area. Its primary limitation is the large number of test items required for any reasonable sized network, as well as the required stability of the item difficulty indices which must be spaced somewhere between .025 and .05 for maximum efficiency. Simulation and empirical studies (this model has received the most extensive empirical assessment) indicate its superior outcome in comparision to conventional testing, where it typically yields high correlations ($r \geq .80$) with a conventional test. The requirement for an exceedingly large pool of test items with known difficulty indices will always be its greatest deterrent.

---

[68] M. Stocking, Short Tailored Tests, Research Bulletin 69-63, (Princeton, N.J.: Educational Testing Service, 1969).

[69] R. L. Linn, D. A. Rock, and T. A. Cleary, "The Development and Evaluation of Several Programmed Testing Methods," Educational and Psychological Measurement 29, (1969): 129-146.

[70] A. G. Bayroff, J. J. Thomas, and A. A. Anderson, Construction of an Experimental Sequential Item Test. Research Memorandum 60-1, (Washington, D.C.: U.S. Army Personnel Research Office, 1960).

[71] See footnote 33 on page 18.

2.3.4 Related Branching Models

If one adjusts the typical branching model for guessing, an appropriate reassignment of the upward and downward movements can be utilized such that upward movements are more minimal than downward movements.[72] This is referred to by Lord as the H-L method.[73] Simulation comparisons indicate that this is a less desirable model than the branching model and has all of its undesirable features, namely, the requirement for a large item pool.

Another variant is the plicate method which adjusts the higher and lower branching steps according to whether the number right is odd or even, or some multiple thereof. Again, it has shown poor performance in comparision to the conventional branching model.[74]

2.3.5 Hybrid Model

This model combines the shrinking and fixed step size methods into a single testing approach. For the first n test items, the difficulty is reduced in a systematic decreasing step manner. Then for the remaining items, a fixed step size is utilized. The main advantage derived lies in its efficiency in establishing an early estimate of a student's performance level using the shrinking step size, and then a refined estimate over very small incremental steps for the remaining items. Again, its primary advantage is improved measurement accuracy in the extremes of the performance continuum, but again it does not resolve the problems for the median range. Therefore, only simulation studies have been performed to date on this model.[75] [76]

2.3.6 Blocked Up and Down Methods

In order to increase the stability of the tests, the blocked up-down model combines several items of approximately equal difficulty into a single block. If a student gets one half of the items correct he is moved up to the next difficulty level block test. If he fails more than half the items, he is moved to an easier block test. Lord investigated

---

[72] A. Birnbaum, "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability," In F. M. Lord and M. R. Novick, eds., Statistical Theory of Mental Test Scores (Reading, Mass.: Addison-Wesley, 1968), Chapters 17 to 20.

[73] See footnote 13 on page 10.

[74] See footnote 13 on page 10.

[75] See footnote 67 on page 24.

[76] See footnote 69 on page 25.

this model utilizing two items per block.[77] Results of the computer simulation indicated that the blocked up and down model is inferior to the single item branching model, although Cleary, Linn, and Rock indicated that a very high correlation coefficient was yielded by using this model on their simulation of test outcomes using actual test scores derived from known ability tests.[78]

2.3.7 Multistage Models

Following the suggestions of Cronbach and Gleser, tests can be constructed so that the initial section provides a routing into three or four performance levels.[79] The second portion of the test, the measurement section, is then administered to each of the subgroups and used to derive a highly accurate score. Rock, Linn, and Cleary evaluated a two-stage routing procedure, a broad range routing procedure, a group discrimination routing procedure, and a sequential item sampling procedure. The group discrimination yielded the most satisfactory results.[80] Subsequent work by Lord indicated that the Robbins Monro process yields the most accurate estimates of a student's performance.[81] This is next followed by the branching model, and the multistage model yields satisfactory results over the full performance range. The biggest problem associated with this model lies in its original construction. Obtaining items of appropriate difficulty and arranging them to achieve the desired result is laborious and voluminous at best. On the other hand, it has obvious applications for a paper and pencil mode.

2.3.8 Flexilevel Model

Created by Lord, the flexilevel model starts a student with a middle difficulty item and proceeds by presenting the next easier item after each wrong response and the next harder item after each correct response.[82] Testing is stopped after n items where n is defined as $(\frac{N}{2} + 1)$ and N is the total number of items of the test. Lord found

---

[77] See footnote 13 on page 10.

[78] See footnote 9 on page 9.

[79] See footnote 12 on page 10.

[80] See footnote 9 on page 9.

[81] See footnote 67 on page 24.

[82] See footnote 6 on page 9.

through computer simulation studies that the flexilevel model yields highly satisfactory results if the difficulty step size is in the range .033 to .067.[83] This model is quite advantageous for two reasons: first, the reduction in test items is clearly specifiable and potential paper and pencil applications are also feasible. Moreover, the test item pool can be directly implemented from an existing conventional test, a highly important developmental factor.

2.3.9 Summary of Tailored Testing

The various problems raised by tailored testing discussed above are summarized by Lord as follows: "Until now, even some very primitive questions about how to carry out tailored testing did not have even vague answers."[84] If these problems are confusing even to the psychometricians, how can the technical training sector have confidence in tailored testing? A mature summary of problems and advantages indicates the wisdom of further research and development.

In some of the studies reported (e.g., Angoff and Huddleston, Cleary, Linn, and Rock), as many as 20% of the students were misclassified by the routing test.[85] [86] [87] In the case of conventional testing, misclassification of students is similarly unavoidable, since no training test of today is perfectly valid and reliable. Given equivalent weakness for each approach, the use of improved test development methodology is the best course of action.

Another serious weakness of tailored testing is that although it is better for the extreme ability groups, it provides less accurate measurement for the average individual than that of a "standard" test.

---

[83] F. M. Lord, "A Theoretical Study of the Measurement Effectiveness of Flexilevel Tests," Educational and Psychological Measurement 31, (1971): 805-813.

[84] See footnote 13 on page 10 (p. 180).

[85] W. H. Angoff and E. M. Huddleston, The Multi-Level Experiment: A Study of a Two-Stage Test System for the College Board Scholastic Aptitude Test, Statistical Report 58-21, (Princeton, N.J.: Educational Testing Service, 1958).

[86] T. A. Cleary, R. L. Linn, and D. A. Rock, "Reproduction of Total Test Score Through the Use of Sequential Programmed Tests," Journal of Educational Measurement 5(3), (1968): 183-187.

[87] See footnote 9 on page 9.

Lord gave tailored testing an apparent "fatal blow" in this comment:

> If, for example, 500 items are available for tailored testing, better measurement will often be obtained by selecting, for example, the n = 60 most discriminating items (highest a) and administering these as a conventional test, rather than by using all 500 in a tailored-testing procedure. This may actually prove to be a fatal objection to any general use of tailored testing.[88]

This remark would hold if tailored testing is applicable only to normative ability measurement, such as the GRE, or the SAT. However, in reaction to this restricted viewpoint of tailored testing, Green argued that "the computer's failure to improve on conventional testing in this situation does not foreclose the possibility of computer advantages in other cases."[89] Very similar opinion was also shared by Crick who reacted: "Lord's restricted view of testing, while certainly a legitimate one, does not exhaust the possible applications of computer-assisted testing.[90]

In discussing the prospects of tailored testing, it seems that the following points are pertinent:

1. One reason for Lord's negative comment on tailored testing is the strategy of comparison with a standard test (i.e., a conventional peaked test). However, in comparing the tailored testing with a "published" (Lord's definition of a conventional unpeaked test) test, his findings indicated that "the tailored procedure gives more accurate measurement than the unpeaked conventional test for all students regardless of level."[91] Thus, in most technical training contexts, tailored testing is apparently the most effective approach.

2. It has also been shown that tailored testing permits a drastic reduction of test items without much loss in the reproducibility of the total test scores.[92] [93] [94]

---

[88] See footnote 13 on page 10 (p. 180).

[89] B. F. Green, "Comments on Tailored Testing," In W. Holzman, ed., Computer-Assisted Instruction, Testing, and Guidance (New York: Harper and Row, 1970) pp. 184-185.

[90] See footnote 4 on page 9 (p. 23).

[91] See footnote 13 on page 10 (p. 179).

[92] See footnote 9 on page 9.

[93] See footnote 86 on page 28.

[94] See footnote 17 on page 11.

3. One novel application was made by Ferguson who used tailored testing in a hierarchical criterion-referenced measurement situation.[95] Concerning the potential usefulness of tailored testing for this purpose, Crick commented: "Intuitively, tailored testing makes much more sense for a criterion-referenced measure than for a norm-referenced measure since the goal of tailored testing is to adjust the test to the individual."[96]

4. In individualized approaches to instruction, it seems that Lord's flexilevel testing may have wide applicability. In the pretest, every subject would take the easy set of the items; but, in the posttest, the subjects would take the difficult set instead. Thus, the use of the parallel forms of the test can be avoided. Furthermore, since the subjects would not have been exposed to many of the harder items, the carry-over effects of testing can be minimized. Although Lord developed the flexilevel testing, he has not emphasized the use of it in this context.

5. Tailored testing is appropriate also in the affective domain of measurement.[97] Tam found that a flexilevel model yielded reliability and validity indices equivalent to the total conventional test, and an empirically observed stop criterion reduced the test length significantly beyond the 50% level.

The prospects of tailored testing depend on willingness to explore its various uses, and the above list is by no means exhaustive. It is hoped that more rigorous explorations of tailored testing will lead to Green's prediction of the "inevitable computer conquest of testing."[98]

2.4 Adaptive Testing for Hierarchical Learning Structures

For the instructional tasks which are hierarchical in nature, special adaptive testing techniques are required due to the known inter-dependencies. The term "hierarchy" is here used in the sense described by Gagné, based on his taxonomy of learning.[99] Gagné proposed that the

---

[95] See footnote 17 on page 11.

[96] See footnote 4 on page 9 (p. 29).

[97] P. T. Tam, "A Multivariate Experimental Study of Three Computerized Adaptive Testing Models for the Measurement of Attitude Toward Teaching Effectiveness" (Unpublished Ph.D. dissertation, Florida State University, 1973).

[98] See footnote 89 on page 29 (p. 194).

[99] R. M. Gagné, "The Acquisition of Knowledge," Psychological Review 69, (1962): 355-365.

prerequisite skills for a terminal objective can be analyzed so that lower ordered skills or behaviors would generate positive transfer to higher level skills. Gagné's method of analysis begins with the terminal objective, and reiterates the following question for each subbehavior (subskill) identified: "What would an individual already have to know how to do in order to learn the new capability simply by being given verbal instructions?"

A number of instances have been reported by both Gagné and Glaser and Nitko in which task analysis procedures have been applied to the study of curricula structures.[100] [101] These include applications dealing with number series, algebraic equations, and elementary geometry.[102] [103] [104] Others include operations with sets, fractions, punctuation, and capitalization of words and reading.[105] [106] [107] Glaser and Nitko further point out that task analysis is a growing area of activity among educational researchers.[108]

---

[100] R. M. Gagné, "Curriculum Research and the Promotion of Learning," Perspectives of Curriculum Evaluation. AERA Monograph Series on Curriculum Evaluation, No. 1, 1967.

[101] R. Glaser and A. Nitko, "Measurement in Learning and Instruction," In R. L. Thorndike, ed. Educational Measurement, (Washington, D.C.: American Council on Education, 1971), 625-670.

[102] See footnote 99 on page 30.

[103] R. M. Gagné and N. E. Paradise, "Abilities and Learning Sets in Knowledge Acquisition," Psychological Monographs 75, No. 14, (1961): (Whole No. 518).

[104] R. M. Gagné and O. C. Bassler, "Study of Retention of Some Topics of Elementary Nonmetric Geometry," Journal of Educational Psychology 54, (1963): 123-131.

[105] W. Hively II, Defining Criterion Behavior for Programmed Instruction in Elementary Mathematics, (Cambridge, Mass.: Committee on Programmed Instruction, Harvard University, 1963).

[106] R. E. Schutz, R. L. Baker, and V. S. Gerlach, Measurement Procedures in Programmed Instruction (Tempe, Arizona: Classroom Learning Laboratory, Arizona State University, 1964).

[107] E. Gibson, "Learning to Read," Science 148, (1965): 1066-1072.

[108] See footnote 101 above.

An initial task analysis results from a rational and subjective procedure, usually performed by curriculum experts. Because of this, it constitutes merely a hypothesized set of relationships involved in the subject matter. If a great deal of faith is going to be put into a hierarchy resulting from a task analysis, empirical validation of the hierarchy becomes necessary. Gagné proposes a simple analysis of data collected on criterion tests referenced to each and all objectives within the hypothesized hierarchy. The subjects on whom data have been collected are first categorized into those who have passed the higher unit and those who passed the lower unit. Implications for the validity of the hierarchy can be made through the simple comparison of the groups.[109]

Applications of techniques for validation purposes include the works of Hively and Schutz, Baker, and Gerlach mentioned earlier, as well as those studies conducted by Gagné and his colleagues.[110] [111] Further applications were demonstrated by Newton and Hickey; Smith and Moore; and Cox and Graham.[112] [113] [114]

Another approach to the validation of task analysis attempts to derive from empirical data statistical indices which can then be used to evaluate the hypothesized hierarchy. These procedures are extensions of the methods of scalogram analysis and simplex analysis.[115] [116] Applications

---

[109] See footnote 100 on page 31.

[110] See footnote 105 on page 31.

[111] See footnote 106 on page 31.

[112] J. M. Newton and A. E. Hickey, "Sequence Effects in Programmed Learning of a Verbal Concept," Journal of Educational Psychology 56, (January 1965): 140-147.

[113] W. I. Smith and J. W. Moore, Learning Sets in Programmed Instruction, Final Report, United States Office of Education Grant No. 7-D-48-0070-208, Lewisburg, Pa.: Bucknell University, 1965.

[114] R. C. Cox and G. T. Graham, "The Development of a Sequentially Scaled Achievement Test," paper presented at the annual meeting, American Educational Research Association, Chicago, 1966.

[115] L. Guttman, "The Basis for Scalogram Analysis," in S. A. Stauffer, ed., Measurement and Prediction (Princeton: Princeton University Press, 1950).

[116] L. Guttman, "A New Approach to Factor Analysis: The Radex," In P. F. Lazarsfeld, ed., Mathematical Thinking in the Social Sciences (Glencoe, Illinois: Free Press, 1954).

have been demonstrated by Resnick; Resnick and Wang; and Boozer and Lindvall.[117] [118] [119] These last researchers concluded that, in general, scalogram analysis seemed more applicable to within-objective hierarchies whereas simplex analysis seemed more appropriate for between-objectives relationships. Okey has presented an overview of the literature on validation of hierarchies.[120]

Hierarchically based instruction, then, has been or can be developed so that tutorial, drill and practice, etc., materials cover each subskill, under the assumption that acquisition of all subskills or necessary behaviors is prerequisite to performing the terminal behavior. That is, based on Gagné's theory of a learning hierarchy, if a student can perform the terminal skill, he is also capable of performing the subordinate skills (or mastered the subordinate skills simultaneously with the terminal objective).

The implications for testing, or for integrating testing with instruction, are in the location of instructional dependencies, and in maximizing the opportunities of the student to participate in training only in those areas in which he does not already have competencies. Pretesting over all objectives in a course can be predicted to be prohibitively lengthy, even though computer capabilities permit total pretesting and then branching to the lowest level objective unmastered, either for finer grained testing, or for instruction and drill, testing for mastery, and movement to the next objective not mastered. Taylor[121] describes a model for integrating testing and instruction which precludes unnecessarily testing on questions which can be assumed, from hierarchical placement, to be unmastered by the learner. In each section of the training sequence,

---

[117] L. B. Resnick, Design of an Early Learning Curriculum, Working Paper 16, (Pittsburgh, Pennsylvania: Learning Research and Developing Center, University of Pittsburgh, 1967).

[118] L. B. Resnick and M. C. Wang, "Approaches to the Validation of Learning Hierarchies," paper presented at the Eighteenth Annual Western Regional Conference on Testing Problems, San Francisco, May, 1969.

[119] R. F. Boozer and C. M. Lindvall, "An Investigation of Selected Procedures for the Development and Evaluation of Hierarchical Curriculum Structures," paper presented at the annual meeting of the American Educational Research Association, New York, February 4-7, 1971.

[120] J. R. Okey, "Developing and Validating Learning Hierarchies," AV Communications Review 21(1), (1973): 87-108.

[121] S. S. Taylor, "Drill and Practice Models," paper presented at the annual Meeting of the American Educational Research Association, New Orleans, La., 25 February-1 March, 1973.

pretest items related to the terminal objective are presented first. The student who answers all items for the terminal objective correctly is branched on to the next section. If he fails one of the terminal objective pretest items, however, he is presented with pretest items for the subordinate skill. If he then fails to respond correctly to all items in this pretest, he is branched to the pretest for the lowest level subordinate skills. If he answers 100% of these items correctly, he is routed to the next level subordinate skills; if he does not answer 100% correctly, he is immediately given instruction on the skills, and is led through a number of drill and practice problems on the skill. The number of items presented for drill varies with the performance of the learner. If he answers 80% or more of the problems correctly on the first attempt, he is moved on to a new topic. Otherwise, he is branched back to the beginning of the instructional sequence. When he displays mastery, he is tested on the next higher skills, and so on. Eventually, he is tested again on the terminal objective.

In a study by Taylor using these techniques, the 300 plus students who, on entering the integrated instruction/testing program, answered all 33 pretest items correctly, were branched past all instructional sequences. Thus they were able to complete the instructional program in approximately 20 minutes, as opposed to several hours of possible testing, drill, and practice for the student in the nonmastery situation.

Ferguson utilized a sequential testing model to move students through an elementary mathematics hierarchy.[122] Grade level was utilized as the entry prediction and placement. The operations were judged to be satisfactory although lack of comparative data prevented an evaluation.

For technical training, the optimal prediction and entry plus a flexible (both upward and downward) movement would be required. In addition, the prediction of potential transfer (synthesis of subskills) would allow for pretesting and training time savings if successful. Student directed decisions might be important in this transfer prediction process due to self-awareness and confidence levels.

2.5 Scoring, Diagnosis, Interpretation, and Reports

For this highly important third process of adaptive testing, limited research findings (theoretical, simulated, or empirical) have been reported. The historical reviews above subsume the preponderance of work today. Therefore, this section will focus on promising topics of further study.

Most scoring procedures utilized the dichotomous right-wrong summed score. Three promising alternatives appear to be feasible. First, one could differentially weight items so that the most discriminating items relative to the criterion decision zone rather than the total score have

---

[122] See footnote 17 on page 11.

the most decisive influence. Studies of item weight indicate weighting can improve decision making as well as test psychometric characteristics.[123] [124] [125] Thus alternative weighted scoring procedures are promising and feasible given a computer's calculation capacity.

In turn, the aggregating or summation process for total score should be studied. Green posits that a mean of difficulty indices for correct responses offers the most accurate procedure.[126] Similar composite score procedures that stress minimally acceptable mastery levels should be investigated.

Finally, there is important information in the error responses elicited from students. Bock proposes an item estimation procedure that yields differential information from error alternatives.[127] Intuitively, a "nearly correct" response is more adaptive than a "dum-dum" response. In turn, these error patterns may yield highly important differential categories of students who have partial knowledge. For one group, the remedial alternative of test item review would be sufficient to achieve mastery while the other extreme group may achieve mastery only through a totally new training strategy. Large student flow and a computer are required to implement the Bock model; fortunately, Air Force technical training satisfies these requirements.

In terms of diagnostic requirements, total test scores and item pass-fail indices are far too summarized for instructional inference making. Measurement in technical training should yield an individual performance profile that indicates the structure and "valley" of weakness. Profile techniques could yield insights like "the verbal indices are so low that only a high multimedia with audio training approach will insure mastery," or "the uniform pattern of indices indicates that incentives to enhance motivation will insure fast mastery." While speculative in nature, the individual performance profiles interface directly into an adaptive instructional model at this operational juncture.[128]

---

[123] J. C. Nunnally, Psychometric Theory (New York: McGraw-Hill, 1967).

[124] J. C. Stanley and M. C. Wang, Differential Weighting (New York: College Entrance Examination Board, 1968).

[125] W. Dick, "Item Weighting: Test Parameter Effects and a Comparison of the Effects of Various Weighting Methods," (Unpublished Ph.D. dissertation, Pennsylvania State University, 1965).

[126] See footnote 89 on page 29.

[127] R. D. Bock, "Estimating Item Parameters and Latent Ability when Responses are Scored in Two or More Nominal Categories," Psychometrika 37, (1972): 29-51.

[128] See footnote 1 on page 6.

Interpretation of adaptive tests can be viewed as a "clinical vs. actuarial" challenge. As sufficient test data bases are collected, refined classification techniques (discriminant analysis) and statistical decision models can be constructed so as to improve the predictive aspects of the interpretation. While a futuristic form of research, the ultimate requirement should be investigated so as to have the full potential of adaptive training (instruction and testing) achieved.

In regard to reports, the recurrent problem of understanding numerical or statistical outputs by instructors, supervisors, etc., will be present. Graphical and verbal reports should be considered and studied. The sufficiency of information for instructional decision making and monitoring is critical. As cited in the Hansen, Hedl, and O'Neil review,[129] automation of the report process is both feasible and desirable in terms of cost and resource utilization. A consumer survey methodology could be profitably employed at this stage. Obviously, adaptive tests will only be useful to the degree that their results are utilized in a sound, rational manner.

---

[129] See footnote 24 on page 12.

III. A DESIGN FOR VALIDATION

3.0 Implementation and Demonstration of Adaptive Testing

This section will describe the project team's experience in implementing and demonstrating the feasibility of adaptive testing paradigms. This involved four major activities, in sequence: analysis of two Air Force technical training courses; identification of course sections appropriate for feasibility/validation study; programming of the computer; and design of appropriate follow-on validation and research studies. After an overview of the first three activities, these follow-on studies will be described in detail.

3.1 Overview

The initiating activity consisted of an extensive literature search relating to all facets of adaptive testing. This eventuated in the literature review presented in the prior section. After appropriate consideration of the state of the art, it was recognized that not all of the fruitful topics raised in the literature search could be implemented given the constraints to be described. Therefore, priorities arranged according to benefits for adaptive testing applications in technical training were delineated and used in the design process. The constraints confining the project in turn delimited the scope and priority structure of the proposed studies.

3.1.1 Priorities. In reference to priorities, it was the project team's judgment that the potential savings in test length, with its concomitant reduction in measurement time, was first and foremost in importance. Secondly, demonstration of the use of computers to improve the accuracy of the testing process, the reduction of instructor involvement, and increase in information for critical decision making was judged essential to establishing the feasibility of a computer-based adaptive testing approach.

As a third priority, the application of adaptive testing to hierarchical learning structures was considered highly important in terms of its potential implications for savings in training time. Fourth, the use of adaptive testing for affect, or course critique activities, was considered feasible within the time constraints, and potentially demonstrative of the breadth of the adaptive testing approach. Finally, the study of the marginal student (the criterion zone decision making problem) was judged to be critical to the technical training mastery learning question. These priorities are ranked according to importance and guided the design efforts in the layout of the proposed three studies.

3.1.2 Constraints. As in all naturalistic situations, numerous constraints shape the nature and scope of research activities. First, the student flow and associated time limitations strongly influenced the size of the endeavors. Secondly, available resources, especially

computer terminals, influenced the team's approach and scope.[130] Finally, the consideration of a strategy least interruptive to the conventional ongoing training affected the nature of the design. In total, while the constraints shaped the initial feasibility study, the demonstration, and subsequently proposed studies, they did not prove to be insurmountable barriers, and they illustrate the manner in which adaptive testing can be readily introduced into ongoing technical training programs.

3.1.3 Demonstration. The variable entry flexilevel testing paradigm, the course critique assessment, and the hierarchical structure paradigm were computer programmed and are currently available on the University of Illinois PLATO system. Given ongoing course revisions, test items are undergoing changes on a frequent basis. However, the basic structure of the tests is coded, and the tests are presently running as will be reviewed under Studies One and Two. As will be described, the team's experience indicated that adaptive testing is an easy measurement process to implement on a large general purpose computer with a viable operating system and training-oriented language. The demonstration was therefore judged highly successful.

3.2 Air Force Course Analysis and Liaison Activity

Concurrent with the literature search for adaptive testing, the project staff performed a task analysis on the Inventory Management (IM) course, and the Precision Measuring Equipment Specialist (PME) course. Given prior task analyses, this process was greatly facilitated.[131] [132] After appropriate consideration of priorities and topics, IM end-of-block III exam and lesson and Block IV exams were selected for demonstration of criterion-oriented adaptive testing. To give the reader some understanding of the structure of this material, the following IM course tasks can be listed:

1. Define equipment item terms.
2. Define Air Force Equipment Management System (AFEMS) abbreviations.
3. Identify AFEMS organization responsibilities.
4. Identify AFEMS chain of command.
5. Obtain higher level approval of request for equipment authorization.
6. Validate at base level request for equipment authorization.
7. Establish Equipment Authorization Inventory Data (EAID)/in-use detail record.
8. Identify procedures for equipment issue.
9. Process equipment turnin.
10. Process intercustody receipt account transfer.

---

[130] Only two to four terminals will be available during FY75 for this activity.

[131] See footnote 1 on page 6.

[132] B. Fallentine, Interim Report: Individualized Inventory Management Course, TM-4775, (Santa Monica, Calif.: System Development Corporation, 1972).

11. Process issue and turn in of vehicles.
12. Issue and turn in nonexpendable organizational items.
13. Issue and turn in nonexpendable personal retention items.
14. Issue and turn in nonexpendable tools.
15. Issue and turn in expendable individual equipment and tools.
16. Identify equipment inventory preparation procedures.
17. Identify procedures/steps required to conduct a physical inventory.
18. Identify procedures for processing a consolidated inventory adjustment document.
19. Identify inventory procedures for vehicles and family quarters.
20. Identify Industrial Plant Equipment (IPE) items and procedures.
21. Identify War Readiness Materials (WRM) procedures.
22. Identify nature and purpose of the EAID/in-use asset report.

For hierarchical learning structure, Block I of PME was selected. This block in essence presents the mathematical concepts and skills necessary for this highly technical, electronics-oriented course. Topics for this course can be subsumed in the following:

1. Applied Mathematics
2. DC Circuit Analysis
3. AC Circuit Analysis
4. Vacuum Tubes and Solid State Principles and Power Supplies
5. Solid State and Vacuum Tube Amplifiers
6. Wave Generating and Shaping Circuits
7. Test Equipment Troubleshooting and Repair Procedures
8. DC and Low Frequency AC Measurement I
9. DC and Low Frequency AC Measurement II

To provide course liaison and cooperative design and demonstration activities, the research team met with course instructors and supervisory personnel from the IM course and the PME course during the third week of July, 1973. For the IM course, arrangements were made to collect the end-of-block test for Blocks III and IV and the criterion progress checks (CPC's) for lessons 1, 2, 8, 9, and 10 of Block IV. The block tests were 50-item multiple choice tests covering about two weeks of instruction. The CPC's consisted of a combination of short answer and form completion items. The CPC's covered a shorter period of instruction than the block tests, and were used mainly to assess performance; for example, form completion. Satisfactory performance on the CPC's is a prerequisite for taking the block examinations. Only those CPC's were selected that were compatible for implementation on the PLATO System; that is, one word or short answer formats. Given the reliability of the PLATO system, graphic displays and natural language processing were not included in this research effort. Logistics of test collection, storage, and security were discussed with Air Force personnel. A procedure was arrived at that interfered least with the normal administration of the course: 1. Block tests and CPC's were turned over to the civilian head of the course by the instructor responsible for coordinating test collection. 2. Tests were stored, and

periodically forwarded to the research team for item analysis. 3. All forms of the tests were collected, as well as the course critique items, for entry into the PLATO system.[133]

For the PME course, similar data collection procedures were arranged. The test items were also entered into the PLATO system. The same course critique items were also planned to be utilized for the PME/ hierarchical learning structure experiment.

3.2.1 Item parameter analysis. In order to perform item parameter analysis, the block test data had to be transferred from the standard Air Force answer sheets to the IBM mark sense sheets. The test data were transferred by assistants and appropriate quality control procedures were followed. In order to carry out the modified flexilevel testing procedures, certain statistical information on the test items was required, i.e., item difficulties and beta weights for predictions. Two types of analysis were performed; item parameter estimates, and stepwise multiple regression. Item analysis of the data was performed on a computer program that allows for the following:

1. Specification of either norm- or criterion-referenced evaluation.
2. Three types of correlations:
   (a) point-biserial correlation of item scores with total score,
   (b) biserial correlation of item scores with total score,
   (c) phi correlation of item scores with test score above or below criterion or median.
3. Selective efficiency--the biserial or point-biserial correlation, between item scores and test scores, corrected for the effect of item difficulty.
4. Reliability:
   (a) norm-referenced (KR-20),
   (b) criterion-referenced.
5. Output:
   (a) the above correlations for correct and incorrect alternatives,
   (b) frequency distribution of raw scores,
   (c) descriptive statistics for each item and all items combined,
   (d) a list of items sequenced according to difficulty and discrimination index,
   (e) a list of students by name or ID number with raw score, percent correct, percentile, and standard score (T-score),
   (f) feedback to each student regarding performance relative to the criterion.

An example of this output is supplied in Appendix A, Computer Program Output.

---

133 Since course critique items were to be flexilevel administered, the standard Air Force format could not be employed in all cases. In the place of the standard Air Force format, four 21-item critiques were developed by the research team. Each of the new course critiques centered on only one aspect of course instruction, such as instructor effectiveness.

In reference to the predictive entry requirement of the test, appropriate stepwise regression analyses were performed on a preliminary group of 100 students. An available "canned" program supplied the following:[134]

1. Summary statistics on each variable.
2. Correlation matrix.
3. At each step,
   (a) analysis of variance table,
   (b) multiple R and R square,
   (c) beta weights,
   (d) partial correlations, and F and tolerance values for variables not in the equation.
4. Summary table for all variables in the equation.

Using the above regression procedure, it was possible to generate a predictive model based on the following data:

1. Class standing
2. AFQT score
3. AQE administrative score
4. AQE general score
5. AQE electrical score
6. AQE mechanical score
7. Block I written score
8. Block II written score
9. Block III written score
10. Block IV written score

Preliminary examples of these relationships are presented in Table 3.

VARIABLE	MULTIPLE R	R SQUARE	RSQ CHANGE	SIMPLE R	R	BETA
W3	.51267	.26283	.26283	.51267	.35165	.30302
W2	.55533	.30840	.04556	.45207	.24190	.20436
W1	.57246	.32771	.01931	.45233	.20523	.22145
AFQT	.58416	.34124	.01353	.17811	-.09315	-.14335
ADM	.58823	.34602	.00478	.24977	.10085	.09463
GEN	.59445	.35338	.00736	.14958	-.10269	-.14708
ELCT	.59878	.35745	.00408	.26590	.08908	.14473
MECH	.59892	.35870	.00125	.06934	-.03127	-.05464
(CONSTANT)					20.69874	

TABLE 3. IM Block IV (W4) Test Score Prediction Summary Table

---

134 N. Norman, D. H. Bent, and C. H. Hull. Statistical Package for the Social Sciences. New York: McGraw-Hill Book Co., 1970.

As sample size grows sufficiently large, more stable estimates of these predictions will be found. In addition, ongoing study of fluctuations in the beta weights, as successive numbers are added, will assess variability and measurement of error.

3.3 Study One--Flexilevel Validation Study

The purpose of this study is to validate the adaptive testing paradigm (predictive entry and tailored item presentation) for both a knowledge-oriented test and a student-evaluated course critique. The primary goal of the paradigm is a significant reduction in testing time. Using a within-subject design (N = 200 or more), each student will be individually entered in the test and given the flexilevel adaptive item movement procedure. After the student completes the adaptive test, all of the remaining items will be presented. Conventional and adaptive forms of the course critique will be given at the ends of Blocks III and IV.

The independent variables will be (a) individualized entry based on regression techniques, using AFQT, clerical, Block I, and Block II scores, and (b) the flexilevel algorithm with its final score, and (c) the adaptive course critique form. The dependent measures will be (a) conventional total scores, (b) item latencies, and (c) total test times. For analysis, correlational techniques should reveal that the relationship of the adaptive score and the total score is greater than .9, and that the predictive entry yields a relationship at .70 or greater with the above two variables.

Utilizing Lord's prediction, there should be an approximate 50% reduction in test items with a 40% reduction in test time for the adaptive test. (Analysis of variance techniques will be used.) For item latencies, three significantly different distributions can be predicted; blocking data at .05 difficulty about the final adaptive score, the very hard items will have longer latency than the expected performance level, which will be longer than the easier items. For the course critique, there will be a high relationship between the two forms, and instructor feedback should be more positive for the adaptive form.

Conventional reliability assessment will be applied to all test forms at both item and form levels.

3.3.1 Computer implementation. The total test paradigm has been programmed in the Tutor language of the PLATO system. From a student point of view, the procedure runs as follows: the student (a) signs on the computer terminal, (b) enters control processing, (c) the system selects the test and entry level for him, and (d) executes the adjusted flexilevel item presentation which will assess his performance. After he has completed the adaptive portion of the test, all remaining items are presented. If he has not achieved mastery based on standard Air Force scoring procedures, he receives offline remediation. If he has demonstrated an acceptable level of performance, the system then decides whether to (a) assign the next flexilevel test, reenter the student in control processing and once again

begin the flexilevel sequence, or (b) sign him off, an option available to the instructor for acceleration. Figure 1 presents a flowchart of a student moving through each of these answers. A more detailed description follows.

In signing on, the student enters his name and the computer executes a security check designed to limit system accessibility and assure test security. Once he has completed the required sign-on activities, the computer system checks his performance record and aptitude profile to determine which of the 10 tests he is ready to take. The system also determines his entry level in the chosen test. Thus, the student is provided the most timely entry test point in terms of his recorded performance, aptitudes, and current in-course status.

Student readiness indices would include previous training activities, courses completed, formal education, and other objective training indices. His aptitude profile might include his test scores on the AFSC, ASVB, and other Air Force standardized aptitude tests. His current instructional status identifies how far along he has gotten in the course. Together these data enable control processing to almost instantaneously compute a predictor equation based on these variables.

Once the predictor equation is determined, the computer system translates it into an appropriate flexilevel test whose difficulty and scope are adjusted to the student's predicted performance. He is therefore provided an evaluation experience individually tailored to his current status. He executes this test on the computer terminal, a useful medium not only because of its rapid response but also because of its transitory display, which augments test security.

The student enters the test at the difficulty level that has been predicted appropriate. If he misses an item, he continues down the difficulty scale until he gets one correct. This establishes his in-test performance base, from which subsequent flexilevel items orginate.

When he has completed the adjusted flexilevel test, the remaining test items are presented and the student responses are evaluated (see Figure 1). Green's scoring procedure will be used to evaluate the flexilevel portion of the test, while the entire test will be evaluated using standard Air Force performance criterion scoring procedures.[135] Thus, for each student a tailored test score and a conventional test score will be available. If full mastery, based on the entire test score, is achieved, the student is provided the opportunity to take the next lesson. If he elects to, he then reenters control processing and begins the same sequence in the next assigned flexilevel test.

In the case of test failure, the student goes offline for course remedial activities keyed to his learning deficiencies. Following remediation, all students reenter control processing and restart the flexilevel testing cycle.

---

[135]See footnote 89 on page 29.

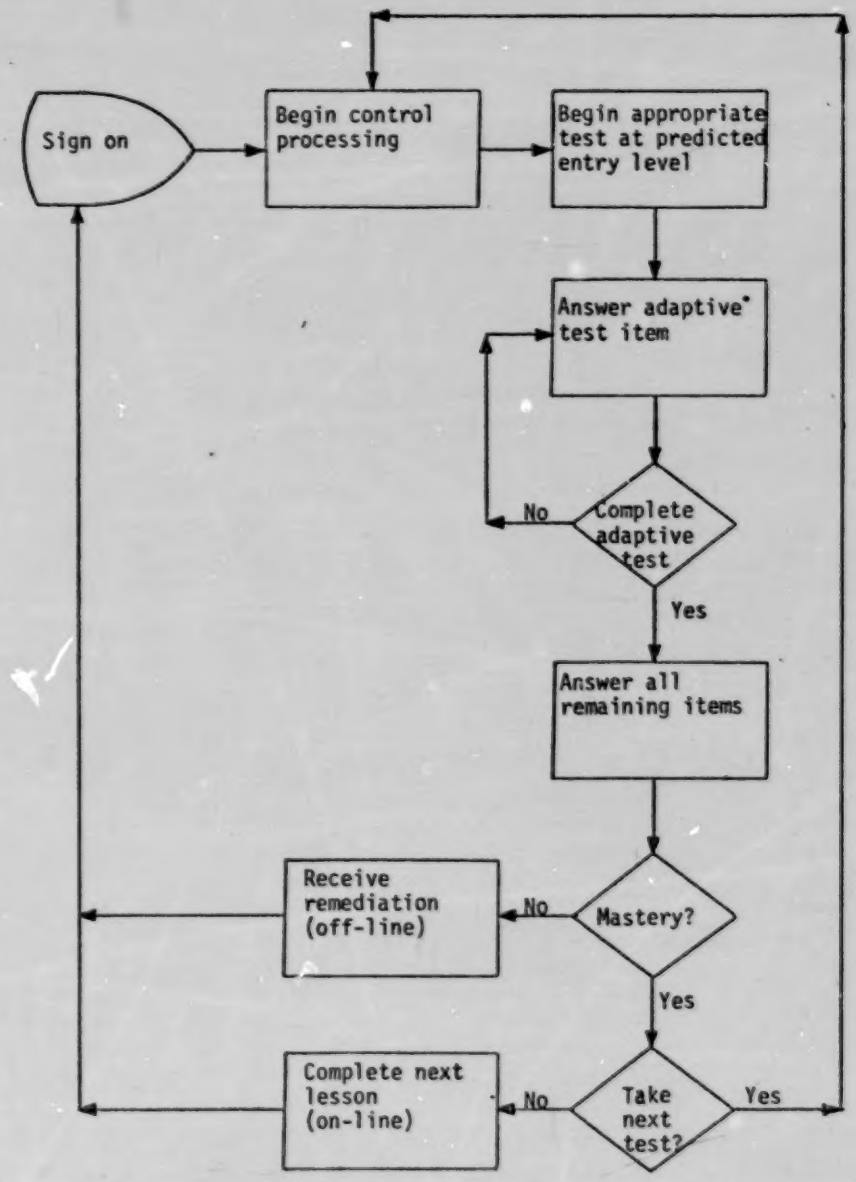

Figure 1. Flowchart of student progress through flexilevel testing program.

After the student attains performance mastery, as a result of either the initial or postremedial 50-item test score, the system then decides if he should continue to the next test. If time permits, he most likely will be routed to control processing for a performance prediction update and subsequent testing. If further testing is not prescribed, he is signed off.[136]

3.4 Study Two--Hierarchical Learning Assessment

The hierarchical learning study will be performed within Block I, Precision Measuring Equipment Specialist (PME) course. Using a within-subject design (N = 100) for validation purposes, an individualized entry prediction based on regression techniques (AFQT, math, mechanical, electronic, and prior math instruction) and a pretest pass-fail unit movement will be the independent variables. All students will be required to attempt all test items with embedded flexilevel branching and subsequent full testing. The dependent measures of total test score, item latencies, and test time will be comparatively analyzed with the adaptive test measures as described in Study One. Besides the reliability analyses, learning time and patterns of unit performance levels will be evaluated. The Gagné pass-fail matrix techniques will be utilized.

3.4.1 Computer testing paradigm. The hierarchical testing paradigm is a strategy designed to minimize testing time while maintaining accurate assessment of the student's level of mastery. Also, it includes prescribing the level of instruction which is most appropriate to the student's real time performance status.

A flowchart of the hierarchical testing paradigm, Figure 2, is included to indicate and clarify the conceptual steps in implementing this strategy. The flowchart illustrates how a student is introduced to the testing paradigm after instruction up to lesson N. An appropriate test is initially administered. If the outcome indicates extreme results, the student's level of mastery is significantly different from what was anticipated; consequently a more appropriate test is administered. When a test which is neither too advanced nor too elementary is administered, the student's response is characterized by moderate performance. At this point, information from all testing and previous performance is summarized to prescribe either (a) remediation, or (b) the next hierarchical lesson, or (c) more advanced instruction. Each of these processes is described below.

1. After the student signs on to the computerized testing system, the system analyzes his aptitudes, performance, and current level of instruction. It then prescribes, using regression-determined prediction equations, the test which is at a level most appropriate for the anticipated performance of the student.

---

[136]Technical documentation for the described flexilevel testing is available to the interested reader on request, by contacting the Air Force Human Resources Laboratory, Lowry Air Force Base, Colorado 80203.

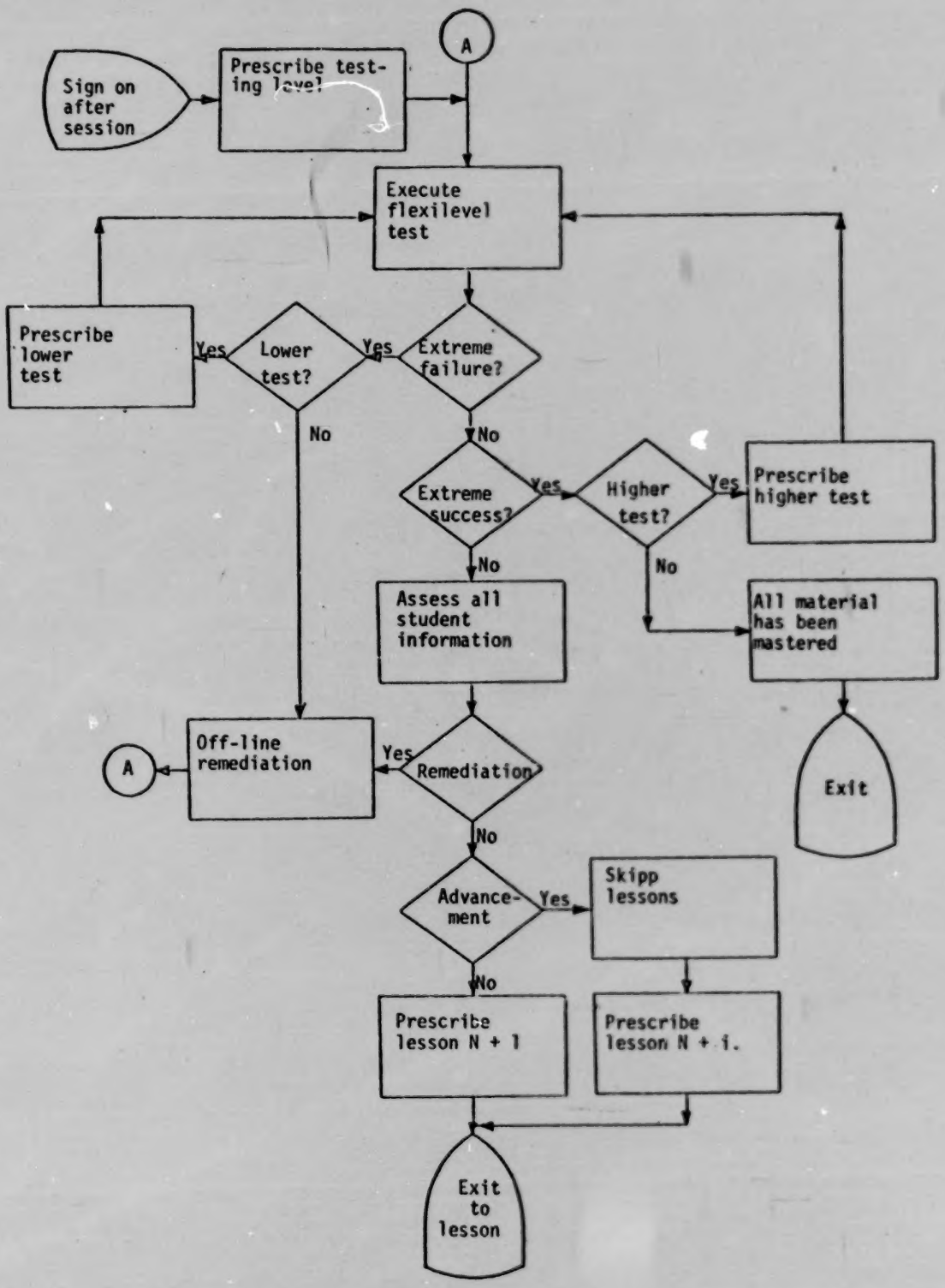

Figure 2. Hierarchical Testing Paradigm

2. A traditional flexilevel midpoint entry test is administered at the predicted level of mastery, and scored using Green's algorithm.

3. An extreme failure indicates that the student's level of mastery is lower than expected.

4. If the student is not at the bottom of the hierarchy, but is lower than predicted, a lower test is available and is prescribed. If the student is at the lowest level in the hierarchy, no lower test will exist, so remediation is administered to raise his level of performance.

5. An extreme success indicates that the student's level of mastery is higher than expected, and a readjustment is in order.

6. If the student is not at the top of the hierarchy, but is higher than predicted, a higher test is available and is prescribed. If the student is at the top of the hierarchy and has had an extreme success, he has mastery of all the material and can be released from further instruction in this area.

7. If the student has done neither extremely poorly nor extremely well, it is clear that his level of mastery is close to that reflected in the test just received. With this information and previous performance data, a decision about appropriate instruction is made with high confidence.

8. If the level of performance is shown to be below what is needed to proceed to the next lesson, remediation is administered and the student's skills are reassessed through the testing procedure.

9. If, on the other hand, the student has demonstrated proficiency at a higher level than that of the next lesson, he is assigned to the most appropriate level (with potential for considerable savings of instruction time).

10. After all is considered, if the student indicates his performance is neither behind nor ahead of the expected level, given his amount of instruction, the next lesson in normal sequence is assigned.

11. With an accurate assessment of the level of mastery, and the appropriate assigned lesson of instruction, the student leaves the testing environment and resumes instruction with the minimum amount of unnecessary testing or instruction.

3.5 Study Three--Criterion Zone Decision Study

Using the result of Studies One and Two, appropriate critical zone criteria will be developed for identifying marginal students. (Subsequent performance, that is, pass-fail patterns, will be used to establish these criteria.) The adaptive computer test will be reorganized so that students scoring in this zone will be randomly assigned to either

(a) A Robbins Monro critical zone sequential test, or (b) analysis of error alternatives using the Bock procedure. Ideally, this study would be performed in both IM and PME. Depending on student flow, at least 50 students in each condition would be compared. Analysis will focus on performance in both the next unit and block. Estimates of time savings will be based on percentage of students in the critical zone and savings on nonwashbacks. Reliability will be based on the performance on the subsequent test.

3.6 Feasibility Results

Since the predominant activity focused on an optimal design and implementation, limited feasibility data or observations can be made concerning the project. Perhaps the most important of these deals with manpower requirements and associated costs. The total effort for the project is approximately one man year. Approximately three-fourths of this man effort was devoted to the study of research, the Air Force course analysis, and the design of the three studies. Approxmately one quarter of the man effort was devoted to computer implementation and data analysis. The important observation is that, given an operational general purpose computer with a modern time-sharing alphanumeric-oriented language, adaptive testing can be implemented in very brief periods of time. The PLATO language was especially well oriented for the preparation of item presentation. The regression equations presented slightly more problems, but not of a significant nature. Two major drawbacks to the PLATO system exist. One is unreliability (at times as much as three days' work effort was lost due to system failure). Secondly, the lack of a general file handling system for storage and retrieval is a problem.[137]

Feasibility in terms of liaison and cooperation with ATC course personnel can be characterized as successful. In the early stages of cooperative information sharing, questions of the appropriateness and test security of this approach raised skepticism. When the instructors were able to see the test items presented on the computer terminal, however, they could more accurately determine the equivalency of the testing method and the capability for time savings for their ongoing instruction. At this stage, the ATC instructional personnel can be characterized as highly cooperative, and interested.

---

[137] University of Illinois personnel are implementing such a file handling system at this time, according to reports given the project team.

IV. CONCLUSIONS AND RECOMMENDATIONS

4.0 Overview

As presented in the prior sections, this project has successfully demonstrated that theoretical models dealing with adaptive testing can be incorporated within the operational measurement requirements of ATC in order to both support ongoing testing, and assess the validity and effectiveness of a computer-based approach for technical training. As a general overview, the design and developmental work to implement adaptive testing proceeded in a most efficient and expeditious fashion. The ready acceptance by ATC instructional personnel and its implications for ongoing operational application speaks to this obvious feasibility. Therefore, the conclusions shall be framed within realization that the adaptive testing models (a flexilevel model for the Inventory Management course and a hierarchical testing model for the Precision Measuring Equipment course) are currently available and can be implemented when existing terminals and implementation support become available. This report now turns to specific conclusions which are framed within the three adaptive testing processes (entry, item tailoring, and scoring/interpretation/reporting).

4.1 Entry Processes

Use of student characteristics (e.g., AFQT) and course performance variables allowed for an individualized variable entry process via linear regression prediction techniques. While the stepwise multiple regression coefficients were moderate in magnitude, the individualized entry should allow for a significant testing time reduction. Moreover, entering each student at his predicted difficulty levels should improve the psychometric characteristics of the process due to the standardization effect and known discrimination effects of presenting test items at the .5 level.

4.2 Tailoring Testing of Item Presentation

The operational feasibility of tailoring items to a student's within test performance was documented by this study. This flexible procedure should allow for time savings of up to 50 percent. The functional interrelation of testing and training within the hierarchical model should yield even more total time savings. The plans for sequentially expanding the criterion test zone should yield even more reliable testing decisions. Moreover, the techniques for analyzing error responses were also implemented in the computer routines. This facet of the study most evidently demonstrates the feasibility and potential of computer-based adaptive testing.

4.3 Scoring, Interpretation, and Reporting

The scoring routines allow for a conventional summed total correct score and an average item difficulty value which can be converted into a

percentile score. Unfortunately, only a limited conventional type student and instructor report were demonstrated. Future developments will undoubtedly indicate the need for verbally-oriented reports.

4.4 Computer Implementation

The University of Illinois Plato System proved to be more than satisfactory for the implementation of the adaptive testing models. As more individualized test composition and reporting procedures are pursued, an improved file handling and report generation capability will have to be available on the computer system. Moreover, improved editing procedures are required if day-to-day revisions are to become operational. Finally, the design coding, debugging, and documentation of the computer-based adaptive testing module with only five man months of effort illustrates the cost-effectiveness of this approach.

4.5 Recommendations

Given the success of this demonstration and feasibility study, the following recommendations appear evident:

1. The empirical validation of the adaptive testing models is paramount. This validation process should take three forms. First, the concurrent and predictive validity of the adaptive testing scores should be related to conventional test scores. Second, the time savings stratified by test type (knowledge, hierarchical, problem solving, etc.) should be analyzed according to cost-effectiveness techniques. Third, the utility of the reports to students and instructors should be assessed to structure any required future extensions and maximize the impact of the measurement process.

2. Future research should focus on the comparative benefits of the flexilevel routines as opposed to Bayesian or Robbins Monro procedures. Those models yielding the greatest joint time savings and amplification of the criterion decision zone should be empirically verified.

3. As the adaptive testing model establishes its empirical validity, a design study of its more dynamic integration into an adaptive instructional model (that is, the role and implication of adaptive testing as both a training strategy and as evaluative feedback to the training system) and its requirements for efficient computer implementation (that is, improved data file handling, editing for revision, and flexible verbal reports) should be pursued.

References

Anastasi, A. "An Empirical Study of the Applicability of Sequential Analysis of Item Selection." Educational and Psychological Measurement 13 (Spring 1953): 3-13.

Angoff, W. H., & Huddleston, E. M. The Multilevel Experiment: A Study of a Two-Stage Test System for the College Board Scholastic Aptitude Test, Statistical Report 58-21. Princeton, N.J.: Educational Testing Service, 1958.

Archambault, F. X. "A Computerized Approach to Scoring Verbal Responses to the Torrance Tests of Creative Thinking." Paper presented at the meeting of the American Educational Research Association, Minneapolis, March, 1970.

Bayroff, A. G. Feasibility of a Programmed Testing Machine. Research Study 64-3. U.S. Army Personnel Research Office, 1964.

Bayroff, A. G., & Seeley, L. C. An Exploratory Study of Branching Tests. Technical Research Note 188. Washington, D.C.: U.S. Army Behavioral Science Laboratory, 1967.

Bayroff, A. G., Thomas, J. J., & Anderson, A. A. Construction of an Experimental Sequential Item Test. Research Memorandum 60-1. Washington, D.C.: U.S. Army Personnel Research Office, 1960.

Birnbaum, A. "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability." In F. M. Lord and M. R. Novick, eds., Statistical Theory of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968, Chapters 17 to 20.

Bock, R. D. "Estimating Item Parameters and Latent Ability When Responses are Scored in Two or More Nominal Categories." Psychometrika 37 (March 1972): 29-51.

Bock, R. D., & Wood, R. "Test Theory." Annual Review of Psychology 22 (1971): 193-223.

Boozer, R. F., & Lindvall, C. M. "An Investigation of Selected Procedures for the Development and Evaluation of Hierarchical Curriculum Structures." Paper presented at the annual meeting of the American Educational Research Association, New York, February 4-7, 1971.

Bryson, R. "Shortening Tests: Effects of Method Used, Length, and Internal Consistency on Correlation with Total Score." Proceedings, 80th Annual Convention of the American Psychological Association, 1972.

Cleary, T. A., Linn, R. L., & Rock, D. A. "An Exploratory Study of Programmed Tests." Educational and Psychological Measurement 18 (Summer 1968): 345-360.

Cleary, T. A., Linn, R. L., & Rock, D. A. "Reproduction of Total Test Score through the Use of Sequential Programmed Tests." Journal of Educational Measurement 5 (Fall 1968): 183-187.

Cogswell, J. F., & Estavan, D. P. Explorations in Computer-Assisted Counseling. TM-2582. Santa Monica, Calif.: System Development Corporation, 1965.

Colby, M. C., Watt, J. B., & Gilbert, J. P. "A Computer Method of Psychotherapy: Preliminary Communication." Journal of Nervous and Mental Disease 142 (February 1966): 148-152.

Cox, R. C., & Graham, G. T. "The Development of a Sequentially Scaled Achievement Test." Paper presented at the annual meeting of the American Educational Research Association, Chicago, Ill., 1966.

Crick, J. E. "A Critical Review of Computer-Assisted Testing." Unpublished qualifying paper, University of Massachusetts, 1972.

Cronbach, L. J. Essentials of Psychological Testing. 3rd ed. New York: Harper and Row, 1970.

Cronbach, L. J., & Gleser, G. C. Psychological Tests and Personnel Decisions. Urbana: University of Illinois Press, 1965.

Dick, W. "Item Weighting: Test Parameter Effects and a Comparison of the Effects of Various Weighting Methods." Unpublished Ph.D. dissertation, Pennsylvania State University, 1965.

Dunn, T. G., Lushene, R. E., & O'Neil, H. F. "A Complete Automation of the Minnesota Multiphasic Personality Inventory and a Study of its Response Latencies." Paper presented at the annual meeting of the American Educational Research Association, New York City, 1971.

Elwood, D. L. "Automation of Psychological Testing." American Psychologist 24 (March 1969): 287-289.

Evans, W. M., & Miller, J. R. "Differential Effects on Response Bias of Computer vs. Conventional Administration of a Social Science Questionnaire: An Exploratory Methodological Experiment." Behavioral Science 14 (May 1969): 216-227.

Fallentine, B. Interim Report: Individualized Inventory Management Course. TM-4775. Santa Monica, Calif.: System Development Corporation, 1972.

Ferguson, R. L. "The Development, Implementation, and Evaluation of a Computer-Assisted Branched Test for a Program of Individually Prescribed Instruction." Unpublished Ph.D. dissertation, University of Pittsburgh, 1969.

Finney, J. C. "Methodological Problems in Programmed Composition of Psychological Test Reports." Behavioral Science 12 (March 1967): 142-152.

Fowler, R. D. "The Current Status of a Computer Interpretation of Psychological Tests." American Journal of Psychiatry 125 (Supp. 1969): 21-27.

Gagné, R. M. "The Acquisition of Knowledge." Psychological Review 69 (July 1962): 355-365.

Gagné, R. M. "Curriculum Research and the Promotion of Learning." Perspectives of Curriculum Evaluation. AERA Monograph Series on Curriculum Evaluation, No. 1, 1967.

Gagné, R. M., & Bassler, O. C. "Study of Retention of Some Topics of Elementary Nonmetric Geometry." Journal of Educational Psychology 54 (June 1963): 123-131.

Gagné, R. M., & Paradise, N. E. "Abilities and Learning Sets in Knowledge Acquisition." Psychological Monographs 75, No. 14, Whole No. 518, 1961.

Gallagher, P. D. An Investigation of Instructional Treatments and Learner Characteristics in a Computer-Managed Instruction Course. Technical Report No. 12. Tallahassee: Florida State University CAI Center, 1970.

Gibson, E. "Learning to Read." Science 148 (May 1965): 1066-1072.

Glaser, R., & Nitko, A. "Measurement in Learning and Instruction." In R. L. Thorndike, ed., Education Measurement. Washington, D.C.: American Council on Education, 1971.

Gleuck, B. C., & Reznikoff, M. "Comparision of Computer-Derived Personality Profile and Projective Psychological Test Findings." American Journal of Psychiatry 121 (June 1965): 1156-1161.

Green, B. F. "Comments on Tailored Testing." In W. Holzman, ed., Computer-Assisted Instruction, Testing, and Guidance. New York: Harper and Row, 1970.

Guttman, L. "The Basis for Scalogram Analysis." In S. A. Stauffer, ed., Measurement and Prediction. Princeton: Princeton University Press, 1950.

Guttman, L. "A New Approach to Factor Analysis: The Radex." In P. F. Lazarsfeld, ed., Mathematical Thinking in the Social Sciences. Glencoe, Illinois: Free Press, 1954.

Hansen, D. N. "An Investigation of Computer-Based Science Testing." In R. C. Atkinson and H. A. Wilson (eds.), Computer-Assisted Instruction: A Book of Readings. New York: Academic Press, 1969.

Hansen, D. N., Hedl, J. J., & O'Neil, H. F. Review of Automated Testing. Technical Memo No. 20. Tallahassee: Florida State University CAI Center, 1971.

Hansen, D. N., Johnson, B. F., Durall, E. P., Lavin, B., & McCune, L. A Rural County Computer-Related Instructional Technology Project. USDHEW Title III Final Report. Tallahassee: Florida State University CAI Center, 1971.

Hansen, D. N., Merrill, P. F., Tennyson, R. D., Thomas, D. B., Kribs, H. D., Taylor, S. T., & James, T. G. The Analysis and Development of an Adaptive Instructional Model(s) for Individualized Technical Training. Technical Report for Contract No. F33615-71-C-1277, Air Force Systems Command. Tallahassee: Florida State University, 1973.

Hansen, D. N., & Schwarz, G. An Investigation of Computer-Based Science Testing. Tallahassee: Florida State University Institute of Human Learning, 1968.

Harding, L. G., Johnson, C. A., & Salop, P. A. An Evaluation of the Use of Chemically Treated Answer Sheets. San Diego, Calif.: Naval Personnel and Training Research Laboratory, 1973.

Hedl, J. J., Jr. An Evaluation of a Computer-Based Intelligence Test. Technical Report 21. Tallahassee: Florida State University CAI Center, 1971.

Hedl, J. J., O'Neil, H. F., & Hansen, D. N. "Computer-Based Intelligence Testing." Paper presented at the annual meeting of the American Educational Research Association, New York City, February, 1971.

Helm, C. E. "Simulation Models for Psychometric Theories." In Proceedings of American Federation of Information Processing Societies 27, Part I. Washington, D.C.: Spartan Books, 1965.

Hively, W. II, Defining Criterion Behavior for Programmed Instruction in Elementary Mathematics. Cambridge, Massachusetts: Harvard University, Committee on Programmed Instruction, 1963.

Kleinmuntz, B. "Personality Test Interpretation by Digital Computer." Science 139 (February 1963): 416-418.

Krathwohl, D. R., & Huyser, R. J. "The Sequential Item Test (SIT)." American Psychologist 11 (August 1956): 419.

Linn, R. L., Rock, D. A., & Cleary, T. A. "The Development and Evaluation of Several Programmed Testing Methods." Educational and Psychological Measurement 29 (Spring 1969): 129-146.

Lord, F. M. "Some Test Theory for Tailored Testing." In W. H. Holtzman, ed., Computer-Assisted Instruction, Testing and Guidance. New York: Harper and Row, 1970.

Lord, F. M. "Robbins-Monro Procedures for Tailored Testing." Educational and Psychological Measurement 31 (Spring 1971): 3-31.

Lord, F. M. "The Self-Scoring Flexilevel Test." Journal of Educational Measurement 9 (Fall 1971): 147-151.

Lord, F. M. "A Theoretical Study of the Measurement Effectiveness of Flexilevel Tests." Educational and Psychological Measurement 31 (Winter 1971): 805-813.

Masang, B. "Item Weighting: An Approach to Invariance of Test Scores under Varying Test Difficulty Levels." Unpublished preliminary paper, Florida State University, 1972.

Massengill, H. E., & Schuford, E. A. Report on the Effect of "Degree of Confidence" in Student Testing. Lexington, Mass.: The Schuford-Massengill Corporation, 1967.

Newton, J. M., & Hickey, A. E. "Sequence Effects in Programmed Learning of a Verbal Concept." Journal of Educational Psychology 56 (January 1965): 140-147.

Norman, N., Bent, D. H., & Hull, C. H. Statistical Package for the Social Sciences. New York: McGraw-Hill Book Co., 1970.

Nunnally, J. C. Psychometric Theory. New York: McGraw-Hill, 1967.

Okey, J. R. "Developing and Validating Learning Hierarchies." AV Communications Review 21 (Spring 1973): 87-108.

Orr, T. B. "A Comparision of the Automated Method and the Face-to-Face Method of Administering the Wechsler Adult Intelligence Scale." Paper presented at the meeting of the Indiana Psychological Association, Indianapolis, April 1969.

Owen, R. J. A Bayesian Approach to Tailored Testing. Research Bulletin 69-72. Princeton, N.J.: Educational Testing Service, 1969.

Paterson, J. J. "An Evaluation of the Sequential Method of Psychological Testing." Unpublished Ph.D. dissertation, Michigan State University, 1962.

Peck, R. F., & Veldman, D. J. An Approach to Psychological Assessment by Computer. Research Memorandum No. 10. Austin: University of Texas, 1961.

Piotrowski, Z. A. "Digital Computer Interpretation of Inkblot Test Data." The Psychiatric Quarterly 38 (January 1964): 1-26.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Denmark Paedogogische Institut, 1969.

Resnick, L. B. Design of an Early Learning Curriculum. Working Paper 16. Pittsburgh, Pennsylvania: University of Pittsburgh, Learning Research and Developing Center, 1967.

Resnick, L. B., & Wang, M. C. "Approaches to the Validation of Learning Hierarchies." Paper presented at the Eighteenth Annual Western Regional Conference on Testing Problems, San Francisco, May 1969.

Robbins, H., & Monro, S. "A Stochastic Approximation Method." The Annals of Mathematical Statistics 22 (1951): 400-407.

Rome, H. P., Swenson, W. M., Mataya, P., McCarthy, E. E., Pearson, J. S., & Keating, R. F. "Symposium on Automation Technics in Personality Assessment." Proceedings of the Mayo Clinic 37 (January 1962): 61-82.

Rosenbach, J. H. "An Analysis of the Application of Utility Theory to the Development of Two-Stage Testing Models." Unpublished Ph.D. dissertation, University of Buffalo, 1961.

Schutz, R. E., Baker, R. L., & Gerlach, V. S. Measurement Procedures in Programmed Instruction. Tempe, Arizona: Arizona State University, Classroom Learning Laboratory, 1964.

Smith, R. E. "Examination by Computer." Behavioral Science 8 (January 1963): 76-79.

Smith, W. I., & Moore, J. W. Learning Sets in Programmed Instruction. Final Report, United States Office of Education Grant No. 7-D-48-0070-208. Lewisburg, Pa.: Bucknell University, 1965.

Stanley, J. C., & Wang, M. C. Differential Weighting. New York: College Entrance Examination Board, 1968.

Starkweather, J. A. "COMPUTEST, a Computer Language of Individualized Testing, Instruction, and Interviewing." Psychological Reports 17 (August 1965): 227-237.

Stocking, M. Short Tailored Tests. Research Bulleting 69-63. Princeton, N. J.: Educational Testing Service, 1969.

Suppes, P., Jerman, M., & Brian, D. Computer-Assisted Instruction: Stanford's 1965-66 Arithmetic Program. New York: Academic Press, 1968.

Tam, P. T. "A Multivariate Experimental Study of Three Computerized Adaptive Testing Models for the Measurement of Attitude toward Teaching Effectiveness." Unpublished Ph.D. dissertation, Florida State University, 1973.

Taylor, S. S. "Drill and Practice Models." Paper presented at the annual meeting of the American Educational Research Association, New Orleans, La. February 25-March 1, 1973.

Veldman, D. J. "Computer-Based Sentence Completion Interviews." Journal of Counseling Psychology 14 (March 1967): 153-157.

Wald, A. Sequential Analysis. New York: Wiley, 1947.

Weiss, D. J., & Betz, N. E. Ability Measurement: Conventional or Adaptive? Research Report 73-1 prepared under Contract No. N00014-67-A-0113-0029 NR No. 150-343, Office of Naval Research. Minneapolis: University of Minnesota, 1973.

Weizenbaum, J. "ELIZA-A Computer Program for the Study of Natural Language Communication between Man and Machine." Communications of the Association for Computing Machinery 9 (June 1966): 36-45.

Wetherill, G. B., & Levitt, H. "Sequential Estimation of Points on a Psychometric Function." British Journal of Mathematical and Statistical Psychology 18 (1965): 1-10.

Wood, R. "The Efficacy of Tailored Testing." Educational Research 11 (1969): 219-222.

Wood, R. "Computerized Adaptive Sequential Testing." Unpublished doctoral dissertation, University of Chicago, 1971.

Wood, R. "Fully Adaptive Sequential Testing: A Bayesian Procedure for Efficient Ability Measurement." Unpublished manuscript, University of Chicago, 1972.

Woods, E. M. "Recent Applications of Computer Technology to School Testing Programs." Review of Educational Research 40 (October 1970): 525-539.

APPENDIX A

COMPUTER PROGRAM OUTPUT

IM Block III Form B Norm-Referenced Analysis Output

IM Block III Form B Criterion-Referenced Analysis Output

IM Block IV Form A Norm-Referenced Analysis Output

IM Block IV Form A Criterion-Referenced Analysis Output

TEST ANALYSIS FOR WALTER DICK A

IMC 303-02

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

PHONE: 599-3660
DATE: 11/13/73
TIME: 10.12.58

IM BLOCK III FORM B
NORM REFERENCED

NUMBER OF STUDENTS = 146
NUMBER OF KEYED ITEMS = 50
NUMBER OF LAST KEYED ITEM = 50

CRITERION SCORE WAS 60.0 PERCENT OF THE ITEMS, OR 30.0 ITEMS CORRECT

PERCENT OF STUDENTS SELECTING THE GIVEN OPTIONS TO EACH ITEM WERE COMPUTED SEPARATELY FOR STUDENTS SCORING ABOVE AND BELOW THE CRITERION SCORE

THE DISCRIMINATORY POWER OF EACH OPTION IS ESTIMATED BY CALCULATING THE POINT BISERIAL CORRELATION BETWEEN STUDENTS SCORES ON THE ITEM AND STUDENT SCORES ON THE TEST
CALCULATING THE INDEX OF SELECTIVE EFFICIENCY (S) WHICH IS THE POINT BISERIAL CORRELATION BETWEEN STUDENT SCORES ON THE ITEM AND STUDENT SCORES ON THE TEST CORRECTED FOR THE EFFECT OF ITEM DIFFICULTY
CALCULATING THE PHI CORRELATION COEFFICIENT BETWEEN STUDENT SCORES ON THE ITEM AND STUDENT PERFORMANCE RELATIVE TO THE CRITERION SCORE

ITEMS RANK-ORDERED ACCORDING TO ITEM DIFFICULTY AND ITEM DISCRIMINATION

FEEDBACK TO INDIVIDUAL STUDENTS RELATIVE TO MASTERY OF COURSE OBJECTIVES WAS NOT REQUESTED

A LISTING OF OBTAINED SCORES BY STUDENT NAMES AND/OR STUDENT NUMBERS IS PROVIDED

PERSONAL HELP IN INTERPRETING ANY PART OF THE FOLLOWING ANALYSIS IS AVAILABLE AT THE OFFICE OF EVALUATION SERVICES, SEMINOLE DINING HALL (599-3128). ASSISTANCE IN THE DEVELOPMENT OF FURTHER EVALUATIVE TECHNIQUES IN ASSESSING AND REACTING TO STUDENT ACHIEVEMENT IN YOUR COURSE IS ALSO AVAILABLE THROUGH THIS OFFICE.

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

MEAN = 39.582
STANDARD DEVIATION = 5.378
RELIABILITY (KR-20) = .768
STANDARD ERROR OF MEAS. = 2.590
RELIABILITY (CRIT.-REF.)= .944 (LIVINGSTON,1972)
INTRACLASS CORR. COEF. = .463 (HAGGARD, 1958)

RAW-SCORE	PERCENT CORRECT	T-SCORE	PERCENTILE	FREQ.	DISTRIBUTION OF RAW SCORES, N = 146
50	100	69	99	1	50 *X
49	98	66	99	2	49 *XX
48	96	66	96	6	48 *XXXXXX
47	94	64	92	5	47 *XXXXX
46	92	62	89	5	46 *XXXXX
45	91	60	85	7	45 *XXXXXXX
44	88	58	79	10	44 *XXXXXXXXXX
43	86	56	72	9	43 *XXXXXXXXX
42	84	54	64	14	42 *XXXXXXXXXXXXXX
41	82	53	55	12	41 *XXXXXXXXXXXX
40	80	51	46	9	40 *XXXXXXXXX
39	78	49	43	7	39 *XXXXXXX
38	75	47	37	11	38 *XXXXXXXXXXX
37	74	45	29	11	37 *XXXXXXXXXXX
36	72	43	24	4	36 *XXXX
35	71	41	21	6	35 *XXXXXX
34	68	40	15	9	34 *XXXXXXXXX
33	66	38	12	2	33 *XX
32	64	36	10	4	32 *XXXX
31	62	34	7	5	31 *XXXXX
30	60	32	4	2	30 *XX
29	58	30	3	1	29 *
28	56	28	3	1	28 *X
27	54	27	3	0	27 *
26	52	25	2	2	26 *XX
25	50	23	1	1	25 *X
24	48	21	0	1	24 *X
23	46	19	0	0	23 *
22	44	17	0	0	22 *
21	42	15	0	0	21 *
20	41	14	0	0	20 *
19	38	12	0	0	19 *

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

(KEYED RESPONSES ARE IDENTIFIED WITH PARENTHESES)

ITEM		A	B	C	(D)	E	OMIT	A	B	C	(D)	E	OMIT			ITEM
				PERCENTS						FREQUENCIES							
1	ABOVE 30.0	2	0	1	96	0	0	3	[illegible]	2	136	0	0	ABOVE 30.0		1
1	BELOW 30.0	0	0	0	100	0	0	0	0	0	5	0	0	BELOW 30.0		1
1	TOTAL GROUP	2	[illegible]	1	97	0	[illegible]	3	0	2	141	0	0	DIFFICULTY =	.97	1
1																1
1	PT BISERIAL	.2[illegible]	[illegible]	-.[illegible]3	.13	0.00	[illegible]							PT BISERIAL DISCRIMINATION =	.13	1
1	S	.74	0.00	-.14	.27	0.00	0.00							S DISCRIMINATION =	.27	1

ITEM		A	(B)	C	D	E	OMIT	A	(B)	C	D	E	OMIT			ITEM
				PERCENTS						FREQUENCIES							
2	ABOVE 3[illegible]	4	94	1	0	[illegible]	1	6	133	1	0	0	1	ABOVE 30.0		2
2	BELOW 30.0	0	100	0	0	0	[illegible]	0	5	0	0	0	[illegible]	BELOW 30.0		2
2	TOTAL GROUP	4	95	1	0	0	1	6	138	1	0	0	1	DIFFICULTY =	.95	2
2																2
2	PT BISERIAL	-.00	-.02	-.02	0.00	0.00	-.04							PT BISERIAL DISCRIMINATION =	-.02	2
2	S	-.[illegible]1	-.[illegible]4	-.14	[illegible].0	0.00	-.23							S DISCRIMINATION =	-.04	2
2	PHI	-.04	-.05	-.02	0.00	0.00	-.02							PHI DISCRIMINATION =	-.05	2

ITEM		(A)	B	C	D	E	OMIT	(A)	B	C	D	E	OMIT			ITEM
				PERCENTS						FREQUENCIES							
3	ABOVE 30.0	76	5	1	18	0	1	107	7	1	25	0	1	ABOVE 30.0		3
3	BELOW 30.0	60	[illegible]	[illegible]	20	[illegible]	2[illegible]	3	0	0	1	0	1	BELOW 30.[illegible]		3
3	TOTAL GROUP	75	5	1	18	0	1	110	7	1	26	0	2	DIFFICULTY =	.75	3
3																3
3	PT BISERIAL	.25	-.09	.10	.25	0.00	.19							PT BISERIAL DISCRIMINATION =	.25	3
3	S	.32	-.24	.63	.41	0.00	.87							S DISCRIMINATION =	.32	3
3	PHI	.07	-.04	-.02	.41	0.00	.30							PHI DISCRIMINATION =	.07	3

ITEM		A	(B)	C	D	E	OMIT	A	(B)	C	D	E	OMIT			ITEM
				PERCENTS						FREQUENCIES							
4	ABOVE 30.0	1	91	4	4	0	0	2	129	5	5	0	0	ABOVE 30.0		4
4	BELOW 30.0	2[illegible]	60	0	0	0	0	1	4	0	0	0	0	BELOW 30.0		4
4	TOTAL GROUP	2	91	3	3	0	0	3	133	5	5	0	0	DIFFICULTY =	.91	4
4																4
4	PT BISERIAL	.24	.32	.16	.15	0.00	0.00							PT BISERIAL DISCRIMINATION =	.32	4
4	S	.91	.52	.50	.48	0.00	0.00							S DISCRIMINATION =	.52	4
4	PHI	.24	.07	-.04	-.04	0.00	0.00							PHI DISCRIMINATION =	.07	4

ITEM		A	B	C	(D)	E	OMIT	A	B	C	(D)	E	OMIT			ITEM
				PERCENTS						FREQUENCIES							
5	ABOVE 30.0	6	2	1	90	0	1	9	3	1	127	0	1	ABOVE 30.0		5
5	BELOW 3[illegible]	40	20	0	40	0	0	2	1	0	2	0	0	BELOW 30.0		5
5	TOTAL GROUP	8	3	1	88	0	1	11	4	1	129	0	1	DIFFICULTY =	.88	5
5																5
5	PT BISERIAL	.25	.12	.06	.29	0.00	.06							PT BISERIAL DISCRIMINATION =	.29	5
5	S	.54	.41	.34	.44	0.00	.34							S DISCRIMINATION =	.44	5
5	PHI	.23	.2[illegible]	-.[illegible]2	.28	0.00	-.[illegible]2							PHI DISCRIMINATION =	.28	5

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

* * * * * * * * * * ITEMS SEQUENCED ACCORDING TO DIFFICULTY INDEX AND DISCRIMINATION INDEX * * * * * * * * * *

ITEM	DIFFICULTY INDEX	ITEM	DISCRIMINATION INDEX
37	1.000	44	.541
12	.993	47	.476
49	.993	31	.466
17	.986	21	.457
19	.986	18	.455
1	.966	13	.432
50	.959	45	.415
38	.952	46	.408
2	.945	24	.400
20	.932	36	.390
42	.932	14	.379
35	.911	35	.374
4	.911	32	.372
7	.911	48	.367
27	.904	30	.355
5	.884	41	.354
47	.884	15	.339
13	.877	22	.328
9	.863	4	.320
10	.856	9	.314
24	.856	25	.311
16	.842	33	.305
8	.829	11	.298
28	.808	5	.293
31	.801	7	.293
18	.795	39	.292
15	.795	27	.277
30	.788	26	.274
40	.788	17	.265
33	.774	23	.255
34	.774	3	.251
22	.767	8	.235
3	.753	28	.227
29	.747	10	.222
45	.726	29	.215
6	.712	43	.215
21	.712	16	.215
26	.692	38	.203
11	.678	50	.196
36	.671	6	.190
23	.664	20	.171
14	.651	34	.162
46	.644	1	.132
44	.616	42	.065
48	.589	49	.055
39	.568	12	.055
43	.514	37	-0.000
41	.507	40	-.003
32	.473	2	-.024

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

NUMBER OF STUDENTS = 146
NUMBER OF ITEMS = 50
MEAN = 39.582
STANDARD DEVIATION = 5.378
RELIABILITY = .768
STANDARD ERROR OF MEAS. = 2.590

LISTING BY STUDENT NAME	I.D. NUMBER	RAW-SCORE	T-SCORE	PERCENTILE	PERCENT CORRECT
	0	28	28	3	56
LMMERT J A K	0	36	43	24	72
ACQUARD JOHN J	0	34	40	15	68
ADAMS PAMELA J	0	45	60	85	90
ADDIN ON DENNIS R	0	35	41	21	70
AGNEW MARK	0	38	47	37	76
ALLEN MONA G	0	36	43	24	72
ASHE THOMAS T	0	34	40	15	68
ATKINS ROBERT M	0	47	64	92	94
BAKER ARNOLD	0	36	43	24	72
BARNEY RANDY R	0	44	58	79	88
BAYLESS MICHAEL L	0	41	53	55	82
BENNET STEPHEN R JR	0	42	54	64	84
BONNELL GARY L	0	47	64	92	94
BRADFORD GEORGE W	0	45	60	85	90
BRIMN DAVID A	0	48	66	96	96
BROOKS EDWARD A	0	43	56	72	86
BROOKS RAYMOND E	0	41	53	55	82
BROWN MARY M	0	35	47	37	76
BUMGARDNER WILLENE M	0	34	40	15	68
BUSIG UDO B	0	38	47	37	76
CAGLE GEORGE A	0	43	56	72	86
CANTU MARIA E	0	37	45	29	74
CANTU PABLO JR	0	40	51	48	80
CAREY DORIS	0	35	41	21	70
CASTER PHILIP N	0	41	53	55	82
CLARK RANDAL G	0	39	49	43	78
CLARK THOMAS J JR	0	40	51	48	80
COLLINS ROSA L	0	47	64	92	94
COOPER JANE F	0	34	40	15	68
CREEHAN THOMAS C	0	43	56	72	86
CROWELL JONATHAN L	0	42	54	64	84
CULBERTSON JOHN H	0	47	64	92	94
CULLIFER NANCY J	0	38	47	37	76
CURTIS REGINALD O	0	41	53	55	82
DAHL GAIL L	0	40	51	48	80
DEHAAN JAMES B	0	24	21	1	48
DESI CE YORS F	0	42	54	64	84
DOMANSKI MARY M	0	44	58	79	88
DRAEGER RICHARD A	0	45	60	85	90
DURHAN MAURICE R	0	45	60	85	90
EATMON JOEY	0	37	45	29	74
ECHERD EDWARD P	0	39	49	43	78
ELSON DAVID A	0	42	54	64	84
FISHER JOHN M	0	25	25	2	52

22/50

TEST ANALYSIS FOR WALTER DICK 9

IMC 343-02

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

PHONE: 599-3660
DATE: 11/13/73
TIME: 10.21.58

NUMBER OF STUDENTS = 146
NUMBER OF KEYED ITEMS = 50
NUMBER OF LAST KEYED ITEM = 50

IM BLOCK III FORM B
CRITERION REF

CRITERION SCORE WAS 6..) PERCENT OF THE ITEMS, OR 30.0 ITEMS CORRECT

PERCENT OF STUDENTS SELECTING THE GIVEN OPTIONS TO EACH ITEM WERE COMPUTED SEPARATELY FOR STUDENTS SCORING ABOVE AND BELOW THE CRITERION SCORE

THE DISCRIMINATORY POWER OF EACH OPTION IS ESTIMATED BY CALCULATING THE BISERIAL CORRELATION BETWEEN STUDENT SCORES ON EACH ITEM AND STUDENT SCORES ON THE TEST
DETERMINING THE DIFFERENCE IN PROPORTIONS OF STUDENTS SCORING ABOVE AND BELOW THE CRITERION SCORE WHO CORRECTLY ANSWERED THE ITEM
CALCULATING THE PHI CORRELATION COEFFICIENT BETWEEN STUDENT SCORES ON THE ITEM AND STUDENT PERFORMANCE RELATIVE TO THE CRITERION SCORE

67 ITEMS RANK-ORDERED ACCORDING TO ITEM DIFFICULTY AND ITEM DISCRIMINATION

FEEDBACK TO INDIVIDUAL STUDENTS RELATIVE TO MASTERY OF COURSE OBJECTIVES WAS NOT REQUESTED

A LISTING OF OBTAINED SCORES BY STUDENT NAMES AND/OR STUDENT NUMBERS IS PROVIDED

PERSONAL HELP IN INTERPRETING ANY PART OF THE FOLLOWING ANALYSIS IS AVAILABLE AT THE OFFICE OF EVALUATION SERVICES, SEMINOLE DINING HALL (599-3128). ASSISTANCE IN THE DEVELOPMENT OF FURTHER EVALUATIVE TECHNIQUES IN ASSESSING AND REACTING TO STUDENT ACHIEVEMENT IN YOUR COURSE IS ALSO AVAILABLE THROUGH THIS OFFICE.

TEST ANALYSIS AND REPORT FOR WALTER DICK B IHC 0J3-12 11/13/73

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

MEAN = 39.459
STANDARD DEVIATION = 5.891

RELIABILITY (KR-20) = .808
STANDARD ERROR OF MEAS. = 2.583

RELIABILITY (CRIT.-REF.)= .946 (LIVINGSTON,1972)
INTRACLASS CORR. COEF. = .583 (HAGGARD, 1958)

RAW-SCORE	PERCENT CORRECT	T-SCORE	PERCENTILE	FREQ.	DISTRIBUTION OF RAW SCORES, N = 146
50	100	68	99	1	*X
49	98	66	99	2	*XX
48	96	64	97	4	*XXXX
47	94	63	93	6	*XXXXXX
46	92	61	89	6	*XXXXXX
45	90	59	84	8	*XXXXXXXX
44	88	58	78	11	*XXXXXXXXXXX
43	86	56	71	9	*XXXXXXXXX
42	84	54	63	13	*XXXXXXXXXXXXX
41	82	53	55	12	*XXXXXXXXXXXX
40	80	51	48	9	*XXXXXXXX
39	78	49	42	7	*XXXXXXX
38	76	48	36	11	*XXXXXXXXXXX
37	74	46	29	10	*XXXXXXXXXX
36	72	44	24	3	*XXX
35	70	42	21	6	*XXXXXX
34	68	41	16	10	*XXXXXXXXXX
33	66	39	12	1	*X
32	64	37	10	4	*XXXX
31	62	36	7	5	*XXXXX
30	60	34	5	2	*XX
29	58	32	4	0	*
28	56	31	4	1	*X
27	54	29	3	0	*
26	52	27	3	2	*XX
25	50	25	2	1	*X
24	48	24	1	1	*X
23	46	22	1	0	*
22	44	20	1	0	*
21	42	19	1	0	*
20	40	17	1	0	*
19	38	15	1	0	*
18	36	14	1	0	*
17	34	12	1	0	*
16	32	10	1	0	*
15	30	8	1	0	*
14	28	7	1	0	*
13	26	5	1	0	*
12	24	3	1	0	*
11	22	2	1	0	*
10	20	0	0	1	*X
9	18	-1	0	0	*
8	16	-2	0	0	*
7	14	-4	0	0	*
6	12	-6	0	0	*
5	10	-7	0	0	*

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

(KEYED RESPONSES ARE IDENTIFIED WITH PARENTHESES)

ITEM		A	B	C	(D)	E	OMIT
				PERCENTS			
1	ABOVE 30.0	1	0	1	96	0	1
1	BELOW 30.0	17	0	0	83	0	0
1	TOTAL GROUP	2	0	1	95	0	1
1							
1	BISERIAL	1.07	0.00	-.14	.44	0.00	-.09
1	U - L	.15	0.00	-.01	.12	0.00	-.01
1	PHI	.21	0.00	-.02	.12	0.00	-.02

A	B	C	(D)	E	OMIT
		FREQUENCIES			
2	0	2	134	0	2
1	0	0	5	0	0
3	0	2	139	0	2

			ITEM
	ABOVE 30.0		1
	BELOW 30.0		1
	DIFFICULTY =	.95	1
			1
BISERIAL DISCRIMINATION =		.44	1
U - L DISCRIMINATION =		.12	1
PHI DISCRIMINATION =		.12	1

ITEM		A	(B)	C	D	E	OMIT
				PERCENTS			
2	ABOVE 30.0	4	95	1	0	0	0
2	BELOW 30.0	0	83	0	0	0	17
2	TOTAL GROUP	4	95	1	0	0	1
2							
2	BISERIAL	-.02	.26	-.07	0.00	0.00	1.32
2	U - L	-.04	.12	-.01	0.00	0.00	.17

A	(B)	C	D	E	OMIT
		FREQUENCIES			
6	133	1	0	0	1
0	5	0	0	0	1
6	138	1	0	0	1

			ITEM
	ABOVE 30.0		2
	BELOW 30.0		2
	DIFFICULTY =	.95	2
			2
BISERIAL DISCRIMINATION =		.26	2
U - L DISCRIMINATION =		.12	2

ITEM		(A)	B	C	D	E	OMIT
				PERCENTS			
3	ABOVE 30.0	76	5	0	19	0	0
3	BELOW 30.0	50	0	17	17	0	17
3	TOTAL GROUP	75	5	1	18	0	1
3							
3	BISERIAL	.34	-.21	1.32	.30	0.00	.60
3	U - L	.26	-.05	.17	-.02	0.00	.17
3	PHI	.12	-.05	.40	-.01	0.00	.40

(A)	B	C	D	E	OMIT
		FREQUENCIES			
107	7	0	26	0	0
3	0	1	1	0	1
110	7	1	27	0	1

			ITEM
	ABOVE 30.0		3
	BELOW 30.0		3
	DIFFICULTY =	.75	3
			3
BISERIAL DISCRIMINATION =		.34	3
U - L DISCRIMINATION =		.26	3
PHI DISCRIMINATION =		.12	3

ITEM		A	(B)	C	D	E	OMIT
				PERCENTS			
4	ABOVE 30.0	1	91	4	4	0	0
4	BELOW 30.0	17	67	0	0	0	17
4	TOTAL GROUP	2	90	3	3	0	1
4							
4	BISERIAL	.64	.66	.38	.36	0.00	1.32
4	U - L	.15	.25	-.04	-.04	0.00	.17
4	PHI	.21	.17	-.04	-.04	0.00	.40

A	(B)	C	D	E	OMIT
		FREQUENCIES			
2	128	5	5	0	0
1	4	0	0	0	1
3	132	5	5	0	1

			ITEM
	ABOVE 30.0		4
	BELOW 30.0		4
	DIFFICULTY =	.90	4
			4
BISERIAL DISCRIMINATION =		.66	4
U - L DISCRIMINATION =		.25	4
PHI DISCRIMINATION =		.17	4

ITEM		A	B	C	(D)	E	OMIT
				PERCENTS			
5	ABOVE 30.0	6	2	1	91	0	0
5	BELOW 30.0	33	17	0	33	0	17
5	TOTAL GROUP	8	3	1	88	0	1
5							
5	BISERIAL	.44	.36	.15	.57	0.00	1.32
5	U - L	.27	.15	-.01	.57	0.00	.17
5	PHI	.26	.18	-.02	.36	0.00	.40

A	B	C	(D)	E	OMIT
		FREQUENCIES			
9	3	1	127	0	0
2	1	0	2	0	1
11	4	1	129	0	1

			ITEM
	ABOVE 30.0		5
	BELOW 30.0		5
	DIFFICULTY =	.88	5
			5
BISERIAL DISCRIMINATION =		.57	5
U - L DISCRIMINATION =		.57	5
PHI DISCRIMINATION =		.36	5

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

* * * * * * * * * * ITEMS SEQUENCED ACCORDING TO DIFFICULTY INDEX AND DISCRIMINATION INDEX * * * * * * * * * *

ITEM	DIFFICULTY INDEX	ITEM	DISCRIMINATION INDEX
49	1.000	12	1.317
37	.993	37	1.317
12	.993	17	.913
17	.979	13	.732
19	.979	31	.686
50	.959	47	.681
1	.952	18	.666
2	.945	44	.661
38	.945	4	.657
20	.925	24	.653
42	.925	21	.625
35	.918	7	.620
7	.911	38	.607
4	.904	27	.584
27	.904	5	.566
5	.884	9	.564
47	.884	45	.562
13	.877	35	.560
9	.863	30	.559
10	.856	14	.522
24	.849	36	.511
16	.842	20	.487
8	.822	48	.471
28	.808	10	.469
31	.801	41	.449
15	.795	32	.448
18	.788	1	.438
30	.781	46	.438
40	.781	8	.427
33	.774	15	.420
22	.767	25	.408
34	.767	26	.401
3	.753	22	.392
29	.740	39	.388
45	.733	28	.386
6	.712	33	.380
21	.705	11	.379
26	.685	23	.366
11	.671	3	.344
36	.671	42	.338
23	.664	50	.333
46	.644	19	.314
14	.644	29	.303
44	.610	6	.289
48	.589	16	.279
39	.568	43	.276
43	.514	2	.258
41	.507	34	.212
32	.473	40	.089

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

NUMBER OF STUDENTS = 146
NUMBER OF ITEMS = 50
MEAN = 39.459
STANDARD DEVIATION = 5.391
RELIABILITY = .838
STANDARD ERROR OF MEAS. = 2.583

LISTING BY STUDENT NAME	I.D. NUMBER	RAW-SCORE	T-SCORE	PERCENTILE	PERCENT CORRECT
	0	28	31	4	56
MMMRSH DOR ENNSS	.	34	41	16	68
RNNG MIER HAROLD E	0	10	0	1	20
ACQUARD JOHN J	0	34	41	16	68
ADAMS PAMELA J	0	45	59	84	90
ADDINGTON DENNIS R		35	42	21	70
AGNEW MARK	0	38	48	36	76
ALLEN MONA G	0	36	44	24	72
ASHE THOMAS T	0	34	41	16	68
ATKINS ROBERT M	0	47	63	93	94
BAKER ARNOLD R	0	36	44	24	72
BARNEY RANDY R	0	44	58	78	88
BAYLESS MICHAEL L	0	41	53	55	82
BENNET STEPHEN R JR	0	44	58	78	88
BONNELL GARY L	.	47	63	93	94
BRADFORD GEORGE W	0	45	59	84	90
BRINN DAVID A	0	48	64	97	96
BROOKS EDWARD A	0	43	56	71	86
BROOKS RAYMOND E		41	53	55	82
BROWN MARY H	0	38	48	36	76
BUMGARDNER WILLENE M	0	34	41	16	68
BUSIG UDO B	0	38	48	36	76
CAGLE GEORGE A	0	43	56	71	86
CANTU MARIA E	0	37	46	29	74
CANTU PABLO JR	0	40	51	48	80
CAREY DORIS	0	35	42	21	70
CASTER PHILIP N	0	41	53	55	82
CLARK RANDAL G	.	39	49	42	78
CLARK THOMAS J JR	0	40	51	48	80
COLLINS ROSA L	0	47	63	93	94
COOPER JANE F	0	34	41	16	68
CREEHAN THOMAS C	.	43	56	71	86
CROWELL JONATHAN L	0	42	54	63	84
CULBERTSON JOHN H	0	47	63	93	94
CULLIFER NANCY J	0	38	48	36	76
CURTIS REGINALD D	.	41	53	55	82
DAHL GAIL L	0	40	51	48	80
DEHAAN JAMES B	0	24	24	1	48
DESINCE YORS F	0	42	54	63	84
DOMANSKI MARY M	0	44	58	78	88
DRAEGER RICHARD A	.	45	59	84	90
DURMAN MAURICE R	0	45	59	84	90
EATMON JOEY	0	37	46	29	74
ECHERD EDWARD P	0	39	49	42	78
EDSON DAVID A	.	42	54	63	84

71/72

TEST ANALYSIS FOR WALTER DICK A

IMC 004-01

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

PHONE: 599-3660
DATE: 11/16/73
TIME: 10.23.37

NUMBER OF STUDENTS = 125
NUMBER OF KEYED ITEMS = 50
NUMBER OF LAST KEYED ITEM = 50

IM BLOCK II FORM A NORM REFERENCED

CRITERION SCORE WAS 60.0 PERCENT OF THE ITEMS, OR 30.0 ITEMS CORRECT

THE DISCRIMINATORY POWER OF EACH OPTION IS ESTIMATED BY CALCULATING THE POINT BISERIAL CORRELATION BETWEEN STUDENTS SCORES ON THE ITEM AND STUDENT SCORES ON THE TEST
CALCULATING THE INDEX OF SELECTIVE EFFICIENCY (S) WHICH IS THE POINT BISERIAL CORRELATION BETWEEN STUDENT SCORES ON THE ITEM AND STUDENT SCORES ON THE TEST CORRECTED FOR THE EFFECT OF ITEM DIFFICULTY
CALCULATING THE PHI CORRELATION COEFFICIENT BETWEEN STUDENT SCORES ON THE ITEM AND STUDENT PERFORMANCE RELATIVE TO THE CRITERION SCORE

ITEMS RANK-ORDERED ACCORDING TO ITEM DIFFICULTY AND ITEM DISCRIMINATION

FEEDBACK TO INDIVIDUAL STUDENTS RELATIVE TO MASTERY OF COURSE OBJECTIVES WAS NOT REQUESTED

A LISTING OF OBTAINED SCORES BY STUDENT NAMES AND/OR STUDENT NUMBERS IS PROVIDED

73

PERSONAL HELP IN INTERPRETING ANY PART OF THE FOLLOWING ANALYSIS IS AVAILABLE AT THE OFFICE OF EVALUATION SERVICES, SEMINOLE DINING HALL (599-3126). ASSISTANCE IN THE DEVELOPMENT OF FURTHER EVALUATIVE TECHNIQUES IN ASSESSING AND REACTING TO STUDENT ACHIEVEMENT IN YOUR COURSE IS ALSO AVAILABLE THROUGH THIS OFFICE.

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

MEAN = 39.080 RELIABILITY (KR-20) = .807
STANDARD DEVIATION = 6.342 STANDARD ERROR OF MEAS. = 2.653

RAW-SCORE	PERCENT CORRECT	T-SCORE	PERCENTILE	FREQ.	DISTRIBUTION OF RAW SCORES, N = 125
50	100	68	99	1	50 *X
49	98	66	99	1	49 *X
48	96	65	98	1	48 *X
47	94	63	95	7	47 *XXXXXXX
46	92	61	96	4	46 *XXXX
45	90	60	86	7	45 *XXXXXXX
44	88	58	8	7	44 *XXXXXXX
43	86	56	73	11	43 *XXXXXXXXXX
42	84	55	67	5	42 *XXXXX
41	82	53	62	8	41 *XXXXXXXX
40	80	52	54	11	40 *XXXXXXXXX
39	78	50	46	10	39 *XXXXXXXXX
38	76	48	36	11	38 *XXXXXXXXX
37	74	47	30	13	37 *XXXXXXXXX
36	72	45	22	8	36 *XXXXXXX
35	70	43	17	5	35 *XXXXX
34	68	42	14	2	34 *XX
33	66	40	12	4	33 *XXXX
32	64	38	10	1	32 *X
31	62	37	7	6	31 *XXXXX
30	60	35	4	1	30 *X
29	58	33	3	2	29 *XX
28	56	32	2	0	28 *
27	54	30	2	0	27 *
26	52	28	2	0	26 *
25	50	27	2	0	25 *
24	48	25	2	0	24 *
23	46	23	2	0	23 *
22	44	22	2	0	22 *
21	42	20	2	0	21 *
20	40	18	2	0	20 *
19	38	17	2	0	19 *
18	36	15	2	0	18 *
17	34	13	2	0	17 *
16	32	12	2	1	16 *X
15	30	10	1	1	15 *X
14	28	8	1	0	14 *
13	26	7	0	1	13 *X
12	24	5	0	0	12 *
11	22	4	0	0	11 *
10	20	2	0	0	10 *
9	18	0	0	0	9 *
8	16	0	0	0	8 *

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

(KEYED RESPONSES ARE IDENTIFIED WITH PARENTHESES)

PERCENTS

ITEM		A	(B)	C	D	E	OMIT
1	TOTAL GROUP	3	92	0	3	4	2
1							
1	PT BISERIAL	.03	.32	0.00	.49	0.00	-.04
1	S	.11	.49	0.00	1.00	0.00	-.16

FREQUENCIES

A	(B)	C	D	E	OMIT			ITEM
4	115	0	4	0	2	DIFFICULTY =	.92	1
								1
						PT BISERIAL DISCRIMINATION =	.32	1
						S DISCRIMINATION =	.49	1

PERCENTS

ITEM		A	B	(C)	D	E	OMIT
2	TOTAL GROUP	0	2	94	3	0	0
2							
2	PT BISERIAL	0.00	.43	.43	.19	0.00	0.00
2	S	0.00	1.00	.67	.67	0.00	0.00
2	PHI	0.00	.53	.48	.19	0.00	0.00

FREQUENCIES

A	B	(C)	D	E	OMIT			ITEM
0	3	118	4	0	0	DIFFICULTY =	.94	2
								2
						PT BISERIAL DISCRIMINATION =	.43	2
						S DISCRIMINATION =	.67	2
						PHI DISCRIMINATION =	.48	2

PERCENTS

ITEM		(A)	B	C	D	E	OMIT
3	TOTAL GROUP	94	4	1	2	0	0
3							
3	PT BISERIAL	.38	.25	.12	.27	0.00	0.00
3	S	.59	.63	.74	1.00	0.00	0.00
3	PHI	.45	.38	-.02	.39	0.00	0.00

FREQUENCIES

(A)	B	C	D	E	OMIT			ITEM
117	5	1	2	0	0	DIFFICULTY =	.94	3
								3
						PT BISERIAL DISCRIMINATION =	.38	3
						S DISCRIMINATION =	.59	3
						PHI DISCRIMINATION =	.45	3

PERCENTS

ITEM		A	B	C	(D)	E	OMIT
4	TOTAL GROUP	2	2	2	74	0	2
4							
4	PT BISERIAL	.16	.16	.33	.25	0.00	.23
4	S	.30	.19	1.00	.34	0.00	.92
4	PHI	-.03	0.00	.23	.16	0.00	.30

FREQUENCIES

A	B	C	(D)	E	OMIT			ITEM
2	25	3	93	0	2	DIFFICULTY =	.74	4
								4
						PT BISERIAL DISCRIMINATION =	.25	4
						S DISCRIMINATION =	.34	4
						PHI DISCRIMINATION =	.16	4

PERCENTS

ITEM		(A)	B	C	D	E	OMIT
5	TOTAL GROUP	93	5	1	2	0	0
5							
5	PT BISERIAL	.36	.06	.15	.52	0.00	0.00
5	S	.55	.20	.92	1.00	0.00	0.00
5	PHI	.42	-.15	.44	.62	0.00	0.00

FREQUENCIES

(A)	B	C	D	E	OMIT			ITEM
116	6	1	2	0	0	DIFFICULTY =	.93	5
								5
						PT BISERIAL DISCRIMINATION =	.36	5
						S DISCRIMINATION =	.55	5
						PHI DISCRIMINATION =	.42	5

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

* * * * * * * * * * ITEMS SEQUENCED ACCORDING TO DIFFICULTY INDEX AND DISCRIMINATION INDEX * * * * * * * * * *

ITEM	DIFFICULTY INDEX	ITEM	DISCRIMINATION INDEX
38	.960	38	.536
40	.952	28	.503
2	.944	35	.488
43	.944	19	.463
3	.936	40	.461
39	.936	33	.449
47	.936	2	.429
37	.928	30	.423
5	.928	16	.419
1	.920	23	.417
6	.920	10	.412
10	.912	36	.406
17	.912	45	.403
7	.896	37	.393
23	.896	24	.389
28	.856	31	.387
11	.840	29	.382
42	.840	41	.381
41	.832	1[illegible]	.378
49	.832	3	.377
22	.824	5	.357
8	.824	22	.354
21	.808	14	.351
24	.8[illegible]	39	.350
27	.800	8	.333
19	.792	21	.329
36	.784	1	.321
31	.784	11	.320
25	.776	25	.315
15	.776	12	.309
35	.760	49	.300
14	.752	7	.295
4	.744	42	.295
34	.736	27	.295
12	.728	15	.287
16	.728	4	.247
45	.696	44	.247
30	.696	6	.238
18	.696	48	.234
32	.680	43	.234
20	.680	13	.212
9	.664	26	.193
33	.640	46	.190
29	.640	32	.174
48	.624	34	.167
26	.616	47	.150
44	.576	9	.141
46	.520	24	.136
50	.480	17	.107

TEST ANALYSIS AND REPORT FOR WALTER DICK A IMC 004-01 11/16/73

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

NUMBER OF STUDENTS	=	125
NUMBER OF ITEMS	=	5.
MEAN	=	39.080
STANDARD DEVIATION	=	6.042
RELIABILITY	=	.807
STANDARD ERROR OF MEAS.	=	2.653

LISTING BY STUDENT NAME	I.D. NUMBER	RAW-SCORE	T-SCORE	PERCENTILE	PERCENT CORRECT
A B EET OMASS P	J	35	43	17	71
AR EL WILLIAM T	J	43	56	73	86
ARI PNN R E K AA	0	33	50	46	78
FAL K N THMO	0	36	48	38	76
JHNSSROBBKTTS	C	40	52	54	80
L K E L LA LEN L	J	47	63	95	94
SH THOMAS TY	0	37	47	30	74
ALLEN WYATT S	0	36	45	22	72
ATKCINS ROBERT M	.	39	50	46	78
B ANCHHBARBARA A	0	31	37	7	62
B OWNNMARY M	0	39	50	46	78
BARNEY BANDYYR	C	42	55	67	84
BAYLESS MICHAEL	J	43	56	73	86
BEN O NOWAIN T	.	4	52	54	80
BENNETT STEPHEN R JR	0	43	56	73	86
BERRY RANDALL L	.	38	48	38	76
BISTLINC JEFFREY D	J	29	33	3	58
BOAGARDNER WILLENE M	.	35	43	17	70
BOINELL GARY L	0	46	61	90	92
BRIHM DAVID A	0	50	68	99	100
BRJHN MYRVELLANE M	C	31	37	7	62
BRZOSTOWSKIIKURT W	.	36	45	22	72
C RROL H RRYYJ	0	34	42	14	68
CAGLE GEORGE A JR	0	39	50	46	78
CAREY DONISI	C	42	55	67	84
CASTER PHILIP	J	38	48	38	76
CHI KKROBERT W	0	43	56	73	86
CLARK THOMAS J JR	0	39	50	46	78
COSTA FRANCES M	J	34	42	14	68
CREEHANNTHOMAS C	0	42	55	67	84
CRJHELL JONATHAN L	.	36	45	22	72
CULLIFER NANCY J	0	41	53	62	82
DAHL GAIL L	C	44	58	80	88
DANIEL PETER F	J	46	61	90	92
DESINCE YVES F	.	44	58	80	88
DRALGER RICHARD A	0	43	56	73	86
DUJAR THOMAS E	0	38	48	38	76
DURHAN MAURICE R	.	35	43	17	70
EATMON JOE T	J	45	60	86	90
ECHERD EDWARD P	C	47	63	95	94
EDSON DAVID A	0	37	47	30	74
ESTRADA JUAN M	0	47	63	95	94
FLAHERTY ANTHONY W	0	31	37	7	62
FLEURRTONNEILEEN M	C	38	48	38	76
FUSELIER JAMES KK		47	63	95	94

77/78

TEST ANALYSIS FOR WHEATER DICK B

IM 004-01

PHONE: 599-3660
DATE: 11/16/73
TIME: 10.24.54

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

IM BLOCK IV FORM A
CRITERION REF

NUMBER OF STUDENTS = 125
NUMBER OF KEYED ITEMS = 50
NUMBER OF LAST KEYED ITEM = 50

CRITERION SCORE WAS 60.0 PERCENT OF THE ITEMS, OR 30.0 ITEMS CORRECT

THE DISCRIMINATORY POWER OF EACH OPTION IS ESTIMATED BY CALCULATING THE BISERIAL CORRELATION BETWEEN STUDENT SCORES ON EACH ITEM AND STUDENT SCORES ON THE TEST
DETERMINING THE DIFFERENCE IN PROPORTIONS OF STUDENTS SCORING ABOVE AND BELOW THE CRITERION SCORE WHO CORRECTLY ANSWERED THE ITEM
CALCULATING THE PHI CORRELATION COEFFICIENT BETWEEN STUDENT SCORES ON THE ITEM AND STUDENT PERFORMANCE RELATIVE TO THE CRITERION SCORE

ITEMS RANK-ORDERED ACCORDING TO ITEM DIFFICULTY AND ITEM DISCRIMINATION

FEEDBACK TO INDIVIDUAL STUDENTS RELATIVE TO MASTERY OF COURSE OBJECTIVES WAS NOT REQUESTED

A LISTING OF OBTAINED SCORES BY STUDENT NAMES AND/OR STUDENT NUMBERS IS PROVIDED

79

PERSONAL HELP IN INTERPRETING ANY PART OF THE FOLLOWING ANALYSIS IS AVAILABLE AT THE OFFICE OF EVALUATION SERVICES, SEMINOLE DINING HALL (599-3128). ASSISTANCE IN THE DEVELOPMENT OF FURTHER EVALUATIVE TECHNIQUES IN ASSESSING AND REACTING TO STUDENT ACHIEVEMENT IN YOUR COURSE IS ALSO AVAILABLE THROUGH THIS OFFICE.

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

MEAN = 38.504 RELIABILITY (KR-20) = .869
STANDARD DEVIATION = 7.319 STANDARD ERROR OF MEAS. = 2.645

RAW-SCORE	PERCENT CORRECT	T-SCORE	PERCENTILE	FREQ.	DISTRIBUTION OF RAW SCORES, N = 125
50	100	66	99	2	50 *XX
49	98	64	98	1	49 *X
48	96	63	97	1	48 *X
47	94	62	94	6	47 *XXXXXX
46	92	61	90	4	46 *XXXX
45	90	59	86	7	45 *XXXXXXX
44	88	57	80	7	44 *XXXXXXX
43	86	56	73	12	43 *XXXXXXXXXXX
42	84	55	66	4	42 *XXXX
41	82	53	61	9	41 *XXXXXXXXX
40	80	52	54	9	40 *XXXXXXXXX
39	78	51	46	10	39 *XXXXXXXXXX
38	76	49	36	10	38 *XXXXXXXXXX
37	74	48	30	10	37 *XXXXXXXXXX
36	72	46	24	6	36 *XXXXXX
35	70	45	20	4	35 *XXXX
34	68	44	17	3	34 *XXX
33	66	42	15	2	33 *XX
32	64	41	13	3	32 *XXX
31	62	40	10	6	31 *XXXXXX
30	60	38	7	1	30 *X
29	58	37	5	3	29 *XXX
28	56	36	4	0	28 *
27	54	34	4	0	27 *
26	52	33	4	0	26 *
25	50	31	4	0	25 *
24	48	30	4	0	24 *
23	46	29	4	0	23 *
22	44	27	4	0	22 *
21	42	26	4	0	21 *
20	40	25	4	0	20 *
19	38	23	4	0	19 *
18	36	22	4	0	18 *
17	34	20	4	0	17 *
16	32	19	4	1	16 *X
15	30	18	3	2	15 *XX
14	28	16	2	0	14 *
13	26	15	1	1	13 *X
12	24	14	1	0	12 *
11	22	12	1	0	11 *
10	20	11	1	0	10 *
9	18	10	1	0	9 *
8	16	8	1	0	8 *
7	-14	7	1	0	7 *
6	12	5	1	0	6 *
5	10	4	1	0	5 *
4	8	3	1	0	4 *
3	6	1	1	0	3 *
2	4	0	0	1	2 *X
1	2	-0	0	0	1 *
0	0	-2	0	0	0 *

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

(KEYED RESPONSES ARE IDENTIFIED WITH PARENTHESES)

ITEM		A	(B)	C	D	E	OMIT
				PERCENTS			
1	TOTAL GROUP	3	91	0	2	0	3
1							
1	BISERIAL	.04	.85	0.00	1.32	0.00	1.06
1	U - L	-.03	.71	0.00	.38	0.00	.37
1	PHI	-.05	.61	0.00	.60	0.00	.51

A	(B)	C	D	E	OMIT			ITEM
		FREQUENCIES						
4	114	0	3	0	4	DIFFICULTY =	.91	1
								1
						BISERIAL DISCRIMINATION =	.85	1
						U - L DISCRIMINATION =	.71	1
						PHI DISCRIMINATION =	.61	1

ITEM		A	B	(C)	D	E	OMIT
				PERCENTS			
2	TOTAL GROUP	.	2	93	3	0	2
2							
2	BISERIAL	0.00	1.09	.86	.41	0.00	1.82
2	U - L	0.00	.24	.46	.18	0.00	.12

A	B	(C)	D	E	OMIT			ITEM
		FREQUENCIES						
0	3	116	4	0	2	DIFFICULTY =	.93	2
								2
						BISERIAL DISCRIMINATION =	.86	2
						U - L DISCRIMINATION =	.46	2

ITEM		(A)	B	C	D	E	OMIT
				PERCENTS			
3	TOTAL GROUP	91	4	1	2	0	2
3							
3	BISERIAL	.89	.55	.32	1.02	0.00	1.52
3	U - L	.57	.22	-.01	.12	0.00	.24
3	PHI	.50	.28	-.02	.23	0.00	.39

(A)	B	C	D	E	OMIT			ITEM
		FREQUENCIES						
114	5	1	2	0	3	DIFFICULTY =	.91	3
								3
						BISERIAL DISCRIMINATION =	.89	3
						U - L DISCRIMINATION =	.57	3
						PHI DISCRIMINATION =	.50	3

ITEM		A	B	C	(D)	E	OMIT
				PERCENTS			
4	TOTAL GROUP	2	20	2	74	0	2
4							
4	BISERIAL	.13	.08	.93	.43	0.00	1.52
4	U - L	-.02	.05	.24	.52	0.00	.24
4	PHI	-.03	.03	.39	.29	0.00	.39

A	B	C	(D)	E	OMIT			ITEM
		FREQUENCIES						
2	25	3	92	0	3	DIFFICULTY =	.74	4
								4
						BISERIAL DISCRIMINATION =	.43	4
						U - L DISCRIMINATION =	.52	4
						PHI DISCRIMINATION =	.29	4

ITEM		(A)	B	C	D	E	OMIT
				PERCENTS			
5	TOTAL GROUP	91	5	1	2	0	2
5							
5	BISERIAL	.80	.02	.46	2.03	0.00	2.53
5	U - L	.57	-.05	.13	.25	0.00	.25
5	PHI	.50	-.06	.34	.49	0.00	.49

(A)	B	C	D	E	OMIT			ITEM
		FREQUENCIES						
114	6	1	2	0	2	DIFFICULTY =	.91	5
								5
						BISERIAL DISCRIMINATION =	.80	5
						U - L DISCRIMINATION =	.57	5
						PHI DISCRIMINATION =	.50	5

OFFICE OF EVALUATION SERVICES
FLORIDA STATE UNIVERSITY

NUMBER OF STUDENTS = 125
NUMBER OF ITEMS = 5.
MEAN = 38.304
STANDARD DEVIATION = 7.515
RELIABILITY = .669
STANDARD ERROR OF MEAS. = 2.645

LISTING BY STUDENT NAME	I.D. NUMBER	RAW-SCORE	T-SCORE	PERCENTILE	PERCENT CORRECT
RR E WILLIAM T	0	32	41	13	64
T M ELLJJEFFE Y	J	15	18	2	30
THIMMESMARTIN G	J	2	0	1	4
ALLEN WYATT S	J	37	48	30	74
ASHE THOMAS TY		39	51	46	78
ATKINS ROBERT M	0	39	51	46	78
B ANDRH A BARA A	0	31	40	10	62
B DAN MARY M	0	39	51	46	78
BACKEY BANDY R		42	55	66	84
BAYLESS MICHAEL	0	43	56	73	86
BEN D N MAIN T	0	40	52	54	80
BENNETT STEPHEN R JR	J	43	56	73	86
BERRY RANDALL L	0	38	49	38	76
BISTLINE JEFFREY D	0	29	37	5	58
BL K ELL LA LEN	0	50	66	99	100
BOMGARDNER MILLENE M	0	35	45	20	70
BONNELL GARY L	0	46	60	90	92
BRINN DAVID A		50	66	99	100
BROWN MYRVELLANE H	0	31	40	10	62
BRZOSTOWSKI KURT W	J	36	46	24	72
CAGLE GEORGE E JR	0	39	51	46	78
CAREY DONALSI		42	55	66	84
CARROLL M KRYYJ	0	34	44	17	68
CASTER PHILIP	0	38	49	38	76
CHICK ROBERT W	0	43	56	73	86
CLARK THOMAS J JR		33	51	46	78
COSTA FRANCES M	J	34	44	17	66
CREEMAN THOMAS C	J	43	56	73	86
CROWELL JONATHAN L	J	36	46	24	72
CULLIFER NANCY J	0	41	53	61	82
DAHL GAIL L		44	57	81	88
DANIEL PETER F	0	46	60	90	92
DESINCE YVES F	0	44	57	80	88
DRAEGER RICHARD A	0	43	56	73	86
DUMAR THOMAS E	0	36	49	38	76
DURMAN MAURICE R	0	35	45	20	70
EATMON JOE T	0	45	59	86	90
ECHERD EDWARD P	0	47	62	94	94
EDSON DAVID A	0	37	48	30	74
ESTRADA JUAN M		47	62	94	94
FLAHERTY ANTHONY W	0	45	59	86	90
FLAHERTY ANTHONY W	0	31	40	10	62
FLEURETON EILEEN M	0	40	52	54	80
FUSELIER JAMES KK		47	62	94	94
GALLEGOS EDWARD K	0	37	48	30	74